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, par convention, 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
0.33 (1/3) et 0.67 (2/3) pour les écarts calibrés.

Les calculs ont été réalisés avec le logiciel SES-Pegase

Dossier CLIMAT (20 points)

Les données suivantes concernent 12 pays européens : Allemagne, Autriche, Belgique, Finlande, Norvège, Royaume-Uni, Suède, Espagne, Italie, Portugal, Hongrie, Pologne.
Pour chacun de ces pays on dispose de deux types d'informations :
1. Les données de l'OMS (Organisation Mondiale de la Santé) qui indiquent, pour l'année 1999, le taux de suicides (pour 100.000 habitants) toutes catégories confondues. Ce taux est noté ici TSTOT.
2. Les données de l'INED (Institut National des Études Démographiques) qui indiquent :
- TEMPT : La température moyenne (en degrés Celsius) sur l'année
- PLUIE : La quantité de précipitations (pluie + neige) en un an (en Litres/m²)

PAYS	TSTOT	TEMPT	PLUIE
ALLE	11.38	8.5	583
AUTR	15.38	9.1	684
BELG	17.30	9.7	833
FINL	21.32	4.5	605
NORV	12.22	4.2	885
ROYA	6.81	10.2	599
SUED	11.78	5.9	569
ESPA	6.49	14.4	492
ITAL	5.88	15.7	828
PORT	3.83	16.0	682
HONG	25.18	10.9	596
POLO	14.24	7.6	511

Sources : http://www5.who.int/mental_health et http://www.ined.fr

Lien entre température moyenne et taux de suicide

Analyses descriptives

On se demande si, dans un pays, le taux de suicide global est lié à la température moyenne du pays. Pour cela on calcule un indice usuellement appelé « corrélation » et noté *r* ou *Rbp*.

1. Indiquer une désignation plus complète de cet indice :

Coefficient de corrélation linéaire de Bravais-Pearson.

2. Cette désignation fait ressortir une caractéristique essentielle de cet indice, qu'il ne faudra pas perdre de vue lors de l'interprétation. Quelle est cette caractéristique ?

Cet indice ne peut détecter que des liaisons linéaires entre les variables

3. En commençant l'analyse par le calcul de cet indice, on passe outre deux recommandations méthodologiques importantes. Quelles sont-elles ?

a. Analyser d'abord chaque variable isolément (l'univarié avant le bivarié !)

b. Analyser le nuage de points (pour détecter les valeurs atypiques, pour voir la forme de la liaison - de type linéaire ou non).

4. On trouve *Rbp* = -.54. Rédiger une conclusion descriptive :

Pour ces 12 pays, la liaison entre la température moyenne annuelle et le taux de suicide est négative : globalement, plus la température moyenne est élevée, plus le taux de suicide est faible. Cette liaison apparaît forte (Rbp = -.54 < -.40)

Inférence

Quelle est la condition préalable essentielle à la mise en œuvre des procédures inférentielles pour envisager une généralisation des résultats à l'ensemble des pays européens ?
(On supposera que cette condition est remplie pour ces données).

Ces pays doivent avoir été choisis au hasard parmi l'ensemble des pays européens

Test T de Student

On trouve t = -2.05 , *ddl* = 10, p = 6.72%.

1. Comment a été calculé *ddl* = 10 :

n - 2 = 12 – 2 = 10

2. Indiquer le résultat du test (significatif / non significatif) en justifiant votre réponse :

Test non significatif
car p > 5% (p = 6.72%)

3. Compte tenu de ce résultat, que peut-on dire de la corrélation parente ?

On ne peut pas rejeter l'hypothèse d'une corrélation parente nulle (entre la température moyenne et le taux de suicide dans l'ensemble des pays européens).
On ne peut pas se prononcer sur l'ampleur de la corrélation parente, ni même sur son signe (+/-)

Intervalle de confiance

1. Quel est l'intérêt de calculer un intervalle de confiance, par rapport à la procédure du test *T* de Student ? On indiquera en particulier les limites d'un test *T* de Student.

Le test T de Student permet uniquement de se prononcer sur l'existence d'une corrélation parente (dans la population) non nulle (lorsqu'il est significatif).
L'intervalle de confiance permet également de se prononcer sur l'ampleur de la corrélation et sur son importance (négligeable, modérée, forte ?)

2. On trouve *IC* (5.00%) = [-0.84; +0.05]. En quoi ce résultat confirme le résultat du test précédent ?

La valeur 0 (corrélation parente nulle) est située dans l'intervalle de confiance. Il s'agit donc d'une valeur de la corrélation parente qu'on ne peut pas rejeter (au seuil 5%) comme nous l'a indiqué le test T *de Student.*

Régression

L'équation de régression linéaire visant à prédire le taux de suicide (TSTOT) selon la température moyenne annuelle (TEMPT) est la suivante :
TSTOT = -0.88 x TEMPT + 21.20

1. Si l'on en croit cette équation :

a. Quel serait l'effet sur le taux de suicide d'une augmentation de 1 degré de la température moyenne d'un pays ?

Si l'on en croit cette équation, une augmentation de 1 degré de la température moyenne d'un pays conduirait à une diminution du taux de suicide (cf. coefficient a négatif) de presque 1 pour 100.000 (plus précisément 0.88 pour 100.000).

b. Quel serait le taux de suicide dans un pays où la température moyenne annuelle est de 0°C ?

Si l'on en croit toujours cette équation, dans un pays où la température moyenne annuelle est de 0°C, le taux de suicide serait de 21.20 pour 100.000 habitants

2. Quel indice permet d'évaluer si cette équation rend bien compte des données observées ?
a. Indiquer son nom complet et le nom abrégé utilisé habituellement pour le désigner

Coefficient de détermination
R^2

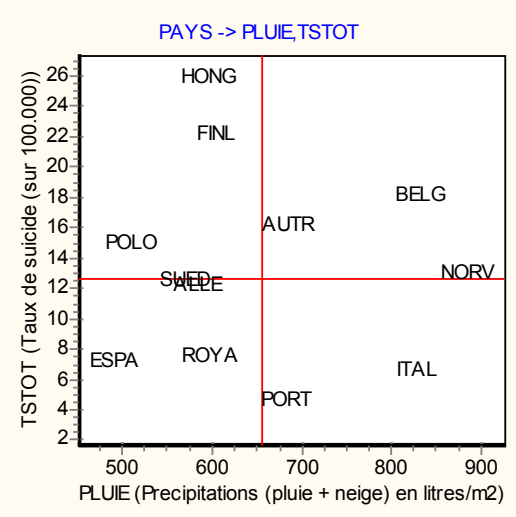
b. Indiquer deux formules de calcul de cet indice :

$R^2 = (r)^2$

$$R^2 = \frac{Var(Yest)}{Var(Y)}$$

Lien entre précipitations et taux de suicide

Analyse du graphe de corrélation



1. Au vu de ce graphique, que peut-on dire du lien entre Précipitation et Taux de suicide dans ces pays ? Justifier votre réponse :

Au vu de ce graphique il semble qu'il n'y ait aucun lien entre ces deux variables (ou, à la rigueur, une liaison négative faible)
En effet, les pays apparaissent dispersés dans les 4 quadrants définis par les axes moyens.

Dossier ALCOOL (6 points)

Ces données sont inspirées d'une étude nationale sur l'alcool et autres drogues, réalisée au Canada en 1989, intitulée « L'alcool au Canada ». On s'intéresse ici au chapitre consacré à la consommation des femmes et particulièrement aux motifs d'alcoolisation mis en avant par les femmes interrogées.

L'enquête distingue 7 catégories d'âge : De 15 à 19 ans, de 20 à 24 ans, de 25 à 34 ans, de 35 à 44 ans, de 45 à 54 ans, de 55 à 64 ans, plus de 65 ans.

Par ailleurs, ces motifs ont été classés selon les 6 catégories suivantes : pour être sociable (SOC), pour agrémenter ses repas (REPA), pour se sentir bien (BIEN), pour se détendre (DETE), pour oublier ses soucis (SOUC), pour être plus à l'aise (AISE).

Source : Site Internet de la santé au Canada.

	SOC	REPA	BIEN	DETE	SOUC	AISE	Total
15A19	64	30	39	29	17	27	206
20A24	76	31	40	39	11	17	214
25A34	74	46	24	37	9	13	203
35A44	73	53	20	36	5	10	197
45A54	73	58	20	31	7	7	196
55A64	73	57	14	24	1	4	173
PLUS65	73	8	16	24	4	11	136
Total	506	283	173	220	54	89	1325

Le tableau représente le nombre de femmes de chaque groupe d'âge ayant mis en avant chacun des motifs. On se demandera si les motifs mis en avant diffèrent selon l'âge des femmes interrogées.

1. Degrés de liberté

Calculer le nombre de degrés de liberté de ce tableau (indiquer la formule générale et le détail des calculs) :

ddl = (J - 1) (K - 1) = (7 – 1) (6 - 1) = 30

2. Effectifs sous indépendance

	SOC	REPA	BIEN	DETE	SOUC	AISE
15A19	78.7	44.0	26.9	34.2	8.4	13.8
20A24	81.7	45.7	27.9	35.5	8.7	14.4
25A34	77.5	43.4	26.5	33.7	8.3	13.6
35A44	75.2	42.1	25.7	32.7	8.0	13.2
45A54	74.8	41.9	25.6	32.5	8.0	13.2
55A64	66.1	37.0	22.6	28.7	7.1	11.6
PLUS65	51.9	29.0	17.8	22.6	5.5	9.1

a. Indiquer le nom plus classique donné à ces effectifs sous indépendance :

On parle plus classiquement des <u>Effectifs théoriques</u>

b. Indiquer comment sont calculés ces effectifs. Pour cela, donner la formule générale puis indiquer comment retrouver la valeur 78.7 (première case en haut et à gauche du tableau) :

$$T = \hat{n}_{jk} = \frac{n_j \times n_k}{n} = \frac{206 \times 506}{1325} = 78.7$$

3. Taux de liaison ou écarts relatifs

TxI	SOC	REPA	BIEN	DETE	SOUC	AISE
15A19	-19 %	-32 %	+45 %	-15 %	+102 %	+95 %
20A24	-7 %	-32 %	+43 %	+10 %	+26 %	+18 %
25A34	-5 %	+6 %	-9 %	+10 %	+9 %	-5 %
35A44	-3 %	+26 %	-22 %	+10 %	-38 %	-24 %
45A54	-2 %	+39 %	-22 %	-5 %	-12 %	-47 %
55A64	+10 %	+54 %	-38 %	-16 %	-86 %	-66 %
PLUS65	**+ 41 %**	**- 72 %**	-10 %	+6 %	-28 %	+20 %

a. On a supprimé les signes des deux taux de liaison en bas et à gauche du tableau (41% et 72%). Ajouter ces signes dans les cases concernées.

b. Commenter le taux de liaison positif le plus élevé (+102%)

Par rapport aux autres groupes d'âge, les femmes les plus jeunes (de 15 à 19 ans) justifient, plus souvent que les autres, leur prise d'alcool par le fait qu'elles oublient ainsi leurs soucis (SOUC).
On en trouve 102% de plus (ou, dit autrement, un peu plus du double) que ce qu'on aurait observé s'il n'y avait pas de différences entre les groupes d'âge du point de vue du motif présenté. En effet elles sont 17 au lieu de 8.4 attendues (et 17 = 8.4 + 102% x 8.4).

Dossier CONDUITE (13 points)

Source : Magazine « Auto-Moto » de décembre 2002.

L'échantillon étudié est constitué de 16 sujets (8 femmes et 8 hommes) tirés au hasard parmi les nombreux candidats à un stage de conduite sportive sur glace.

Tous les sujets sont titulaires du permis de conduire « voiture » depuis au moins 2 ans, mais la moitié des sujets de chaque sexe est également titulaire du permis « moto ». On distingue ainsi deux niveaux de compétence (COMP) : les sujets « mono-compétents » (M) uniquement titulaires du permis « voiture » et les sujets « poly-compétents » (P) titulaires des deux permis.

On teste les réflexes de freinage en présence d'un obstacle. On note le nombre de virages réussis sur l'ensemble du parcours, sans avoir percuté l'obstacle.

s01	s02	s03	s04	s05	s06	s07	s08	s09	s10	s11	s12	s13	s14	s15	s16
P	P	P	P	M	M	M	M	P	P	P	P	M	M	M	M
40	39	34	37	32	29	35	25	26	31	35	33	24	24	19	18

Effet de la Compétence sur la performance totale

On considèrera pour toutes les analyses qui suivent que la différence entre les moyennes est importante si elle est au moins égale à 5 (virages).

Analyse descriptive

On compare tout d'abord les moyennes des deux groupes. On trouve, pour les mono-compétents, une moyenne de 25.8 virages réussis et, pour les poly-compétents, une moyenne 34.4 virages réussis, soit une différence de moyennes, d = 8.6.

1. Pour affiner la mesure de l'ampleur des différences entre les deux groupes on calcule un écart calibré. On calcule la valeur du « *d de Cohen* » noté *d/s*. On trouve *d/s* = 1.6.

a. Indiquer quel est l'intérêt d'un écart calibré (en général) par rapport à la différence brute des moyennes :

Pour comparer des groupes, un écart calibré prend en compte la dispersion à l'intérieur des groupes, et pas seulement l'écart entre les moyennes des groupes.
Avec un écart calibré, pour une certaine différence de moyennes, l'écart entre deux groupes sera déclaré d'autant plus important que la dispersion intra-groupes est faible.

b. Rédiger une conclusion descriptive :

Pour cet échantillon de 16 conducteurs candidats à un stage de conduite sur glace, le groupe des conducteurs poly-compétents (permis auto et moto) a de meilleures performances que le groupe des mono-compétents (permis auto seul) du point de vue du nombre de virages réussis. La différence est importante, que l'on considère l'écart brut entre les moyennes (d = 8.6 > 5) ou l'écart calibré (d/s = 1.6 > 0.67).

Analyse inférentielle

On met en œuvre un test t de Student. On trouve $t[14] = 3.26$ et $p = 0.57\%$.
On calcule également un intervalle de confiance sur la différence des moyennes.
On trouve *IC* (5.00%) = [2.92 ; 14.3]
A partir de ces éléments, rédiger une conclusion inférentielle détaillée qui prolongera la conclusion descriptive ci-dessus :

Il semble que l'on puisse conclure que, pour l'ensemble des conducteurs cette fois, les poly-compétents sont meilleurs que les mono-compétents (t[14] = 3.26, p = 0.57% < 5%). Toutefois on ne peut pas conclure à une différence importante (> 5) entre ces deux groupes (IC (5%) = 2.92 ; 14.3]).

Décomposition de la variance

Lors de l'étape descriptive on aurait également pu calculer un *Eta²*

a. Indiquer deux autres termes ou expressions utilisés pour désigner cet indice :

Rapport de corrélation

Proportion (ou pourcentage) de Variance (PV)

b. Indiquer sa formule de calcul :

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

c. Sur quelle propriété s'appuie le calcul de cet indice ? Indiquer son nom ET la propriété :

Il s'agit de la « Propriété de décomposition de la variance » :

Vtotale = Vinter + Vintra

Calculs (4 points)

Soit les 5 valeurs numériques suivantes : 4, 1, 0, 5, 0

1. Calculer le mode, la médiane et la moyenne de ces 5 valeurs :

   a. Mode = 0

   b. Médiane = 1

   c. Moyenne = 2

2. Calculer la variance de ces valeurs (indiquer le détail des calculs) :

   Var = (4-2)²+(1-2)²+(0-2)²+(5-2)²+(0-2)² / 5 = 4 + 1+ 4 + 9 + 4 / 5 = 22 / 5 = 4.4

Propriétés affines (3 points)

Soit deux variables numériques *X* et *Y* pour lesquelles on dispose des moyennes, écart-type et corrélation :

Moy (X) = 2.5	*Moy (Y)* = 4	*Ety (X)* = 1.5	*Ety (Y)* = 3	R_{BP} *(X,Y)* = -.60

On construit de nouvelles variables (*X'* et *Y'*) dérivées des précédentes par les procédures suivantes : *X'* = *X* + 2 et *Y'* = -3 *Y*.

Calculer et reporter ci-dessous les statistiques sur les nouvelles variables :

Moy (X')	=	4.5	*Moy (Y')*	=	-12
Ety (X')	=	1.5	*Ety (Y')*	=	9
R_{BP} *(X', Y')*	=	+.60			

Combinatoire (6 points)

1. Permutations

Lorsque *n* = 3 personnes (abc) se présentent à l'entrée d'une salle, elles peuvent y entrer dans 6 ordres différents : abc, acb, bac, bca, cab, cba.

a. Quelle formule permet de retrouver ce nombre d'ordres ou « permutations » possibles (*P*) de *n* objets ? Indiquer la formule générale, développer le calcul pour *n* = 3 et retrouver le résultat ci-dessus (6 permutations) :

P = n ! = 3 ! = 3 x 2 x 1 = 6

b. A partir de cette formule, calculer le nombre de permutations possibles pour 5 personnes :

P = n ! = 5 ! = 5 x 4 x 3 x 2 x 1 = 120

2. Combinaisons

Les procédures inférentielles s'appuient sur des procédures d'échantillonnage où on construit tous les échantillons possibles de taille *n* pouvant être construits à partir de la population de taille *N*.

Le nombre d'échantillons possibles s'obtient par la formule ci-dessous. Utiliser cette formule pour calculer le nombre d'échantillons de taille *n* = 2 pouvant être constitués à partir d'une population de taille *N* = 6 (indiquer le détail des calculs) :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{6!}{2! \times 4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) \times (4 \times 3 \times 2 \times 1)} = \frac{6 \times 5}{2} = \frac{30}{2} = 15$$

De la description à l'inférence (5 points)

D'après un exemple de Beaufils (1996). *Statistiques appliquées à la psychologie , tome 2 – Statistiques inférentielles,* Paris : Bréal.

Les résultats à un examen de statistique sont analysés sur un échantillon de 100 sujets tirés au hasard parmi l'ensemble des 590 étudiants ayant passé l'épreuve. Pour chaque étudiant on regroupe d'une part les résultats aux questions portant sur la description (DESC), d'autre part les résultats aux questions portant sur l'inférence (INFE).
Pour chacun de ces sujets et pour chaque partie on note si son score est au moins égal à la moyenne (Réussite) ou Non (Échec). On fait l'hypothèse que les sujets ont mieux réussi la partie description (DESC).

Les données sont présentées sous la forme suivante :

DESC	INFE	Effectif
Réussite	Réussite	45
Réussite	Échec	18
Échec	Échec	12
Échec	Réussite	25

1. Représenter ces résultats sous la forme d'un tableau de contingence, croisant les deux variables binaires DESC et INFE :

		INFE		
		Réussite	Échec	Total
DESC	Réussite	45	18	63
	Échec	25	12	37
	Total	70	30	100

2. Indiquer la particularité de ce tableau de contingence :

Il s'agit de <u>mesures répétées</u> sur une variable binaire

3. Pour l'analyse descriptive des données calculer la différence de pourcentages de réussite entre les deux parties (DESC et INFE) :

d = 63 / 100 – 70 / 100 = -7 / 100 = 7 points de pourcentage

4. Quelle est la partie la mieux réussie (DESC ou INFE) ?

La partie INFE est mieux réussie

5. Quel est le nom du test statistique mis en œuvre classiquement pour tester l'hypothèse nulle d'une absence de différences, dans la population, entre les résultats aux deux parties de l'examen ?

Test Khi² de McNémar

Formules (3 points)

1. Pour évaluer l'ampleur de la liaison globale entre <u>deux variables numériques</u>, le coefficient de Corrélation est un indice dérivé de la Covariance par la formule suivante : $\frac{Cov(X,Y)}{Ety(X) \times Ety(Y)}$.

Cette formule fait ressortir que cet indice peut être vu comme un indice dérivé de la Covariance. Ces deux indices mesurant fondamentalement la même chose, quel est l'intérêt de calculer le coefficient de Corrélation plutôt que la Covariance ou, dit autrement, quel est l'inconvénient de la Covariance ?

La covariance a des <u>bornes (minimum et maximum) qui varient selon les données</u> (elles sont fonction des écarts-type des deux variables). La valeur d'une covariance est donc <u>difficile à interpréter</u>.
<u>La corrélation a des bornes fixes</u> (-1 et +1)

2. Dans le contexte de l'analyse de la liaison entre <u>deux variables nominales (*J* et *K*)</u>, il existe deux indices qui permettent d'évaluer l'ampleur de la liaison globale entre les deux variables et dont l'un est dérivé de l'autre sur le même principe.

a. Quels sont ces deux indices ?

Phi² et V² de Cramer

b. Donner la formule de calcul qui permet de passer de l'un à l'autre :

<u>V² = Phi² / Phi²max</u> (avec Phi²max = Min (J et K) – 1)