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 80 ; la note finale sera donnée sur 20.
Indiquer les réponses exclusivement sur ce document.
Ne rien écrire dans les marges.

NB : On prendra, par convention, les valeurs repères suivantes pour se prononcer sur l'importance d'une corrélation :
0.20 (valeur limite d'une corrélation faible)
0.40 (valeur limite d'une corrélation importante).

Les calculs ont été réalisés avec le logiciel SES-Pegase*.*

Dossier DEPLIANTS (20 points)

Une étude de Geller, Witmer et Orebaugh (1976; d'après Howell, 1998, p.169-173) visait à étudier la tendance à jeter des dépliants sur la voie publique. Ces auteurs voulaient savoir s'il serait efficace d'inclure le message "Ne pas jeter sur la voie publique" dans les dépliants distribués au supermarché.

Des dépliants publicitaires sont distribués aux clients d'un supermarché. Dans la condition *Message,* les dépliants contenaient en plus le message "Ne pas jeter sur la voie publique. Veuillez utiliser les poubelles". Dans la condition *Témoin,* les dépliants ne contenaient que la liste des promotions du jour.

Les auteurs ont relevé le nombre de dépliants a/ jetés dans les caddies, sur le sol… (Jetés)
b/ trouvés dans les poubelles ou non retrouvés, supposés emportés par les clients (Emportés).
Les effectifs observés figurent dans le tableau ci-dessous :

	Jetés	Emportés	Total
Message	290	579	869
Témoin	385	518	903
Total	675	1097	1772

On s'intéresse aux dépliants jetés (Jetés), en faisant l'hypothèse que le message contribuera à leur diminution.

1. Pourquoi ne peut-on pas se contenter de comparer les 2 effectifs de la colonne Jetés (290 et 385) pour répondre à la question ?

Car l'effectif de la condition Message (869) est différent de celui de la condition Témoin (903).

2. On a calculé *P*(Jetés / Message) = 33% et *P*(Jetés / Témoin) = 43%. Indiquer quelles opérations effectuer pour retrouver ces deux pourcentages avec une calculette :

P (Jetés / Message) = 290 / 869
P (Jetés/ Témoin) = 385 / 903

3. On trouve également $Phi^2 = 0.01$. Indiquer la formule qui permet d'en déduire la valeur de la statistique Khi^2 puis calculer cette valeur :

$$Khi^2 = n \times Phi^2 = 1772 \times 0.01 = 17.72$$

4. Indiquer la formule de calcul des degrés de liberté du tableau puis calculer cette valeur :

$$ddl = (J-1) \times (K-1) = 1$$

5. Lors du test de l'hypothèse nulle classique pour un tel tableau, un logiciel nous indique p = 0.006 %. Indiquer en quoi consiste cette hypothèse nulle :

H0 : Dans la population parente… le pourcentage de dépliants jetés est indépendant de la condition (Message ou Témoin) / Il n'existe pas de lien enter la présence du message et la tendance à jeter les dépliants sur la voie publique.

6. Utiliser l'ensemble des résultats rapportés ci-dessus pour rédiger une conclusion complète (descriptive puis inférentielle) qui résume les résultats obtenus.

a. Conclusion descriptive :

Conformément à l'hypothèse, parmi les 1772 dépliants distribués dans ce supermarché, le pourcentage de dépliants jetés est plus faible lorsqu'ils contiennent un message invitant à ne pas les jeter sur la voie publique (33%) que lorsqu'ils ne contiennent que la liste des promotions du jour (43%). Cependant, la différence entre ces deux conditions peut être considérée comme faible ($V^2 = Phi^2 = 0.01 < 0.04$).

b. Conclusion inférentielle :

Il semble que, conformément à l'hypothèse, pour les clients de ce supermarché, le fait d'inclure dans les dépliants un message invitant à ne pas les jeter sur la voie publique, tend à réduire le pourcentage de dépliants jetés ($Khi^2 = 17.72$, $ddl = 1$, $p = 0.006\%$).

7. Pour analyser plus précisément les différences entre les deux groupes, on calcule d'autres indices.
Pour chaque indice indiquer a/ la formule qui a permis d'obtenir cette valeur et b/ le calcul à effectuer (avec les valeurs numériques) pour retrouver la valeur obtenue pour la case en haut à gauche du tableau (Message, Jetés).

a. Effectif sous indépendance : 331

$$\hat{n}_{jk} = \frac{n_j \times n_k}{n} = \frac{675 \times 869}{1772}$$

b. Pourcentage sous indépendance : 19%

$$\hat{p}_{jk} = \frac{\hat{n}_{jk}}{n} = \frac{331}{1772}$$

c. Écart brut à l'indépendance : -41

$$e_{jk} = n_{jk} - \hat{n}_{jk} = 290 - 331$$

d. Écart relatif à l'indépendance (Taux de liaison) : -12%

$$Txl_{jk} = \frac{n_{jk} - \hat{n}_{jk}}{\hat{n}_{jk}} = \frac{O - T}{T} = \frac{290 - 331}{331} = \frac{-41}{331}$$

e. Contribution absolue au Phi^2 : 0.0029

$$Cta_{jk} = \hat{p}_{jk} \times \left(Txl_{jk}\right)^2 = 0.19 \times (-0.12)^2$$

f. Contribution relative au Phi^2 : 32%

$$Ctr_{jk} = \frac{Cta_{jk}}{Phi^2} = \frac{0.0029}{0.01} = 0.32$$

8. On peut également, sur un tel tableau, calculer un indice nommé *Rapport des chances* ou *Odds Ratio*. Cet indice s'obtient en calculant tout d'abord deux rapports (un pour chacun des 2 groupes). Indiquer les calculs à faire pour obtenir ces deux rapports, nommés *Odds* :

a. Pour le groupe « Message » :
Odds = 290 / 579 ou 579 / 290

b. Pour le groupe « Témoin » :
Odds = 385 / 518 ou 518 / 385

Dossier MATERNELLE (20 points)

Dans une école d'une ville anglaise, on souhaite comparer la maturité sociale d'enfants de 7 ans, selon le nombre d'années de préscolarisation effectuées (ou non) en maternelle :
- g0 : pas de préscolarisation en maternelle
- g1 : 1 an de préscolarisation
- g2 : 2 ans, ou plus, de préscolarisation.

On constitue un échantillon de 15 enfants en tirant au hasard 5 enfants dans chacun des 3 groupes.
Chaque enfant a passé une épreuve de maturité sociale notée sur 20. Les données obtenues sont les suivantes :

ENFANTS	MATURITE	PRESCOL
e01	4	g0
e02	8	g0
e03	11	g0
e04	5	g0
e05	3	g0

ENFANTS	MATURITE	PRESCOL
e06	9	g1
e07	14	g1
e08	15	g1
e09	10	g1
e10	12	g1

ENFANTS	MATURITE	PRESCOL
e11	15	g2
e12	18	g2
e13	20	g2
e14	10	g2
e15	14	g2

Source: D'après Hays, W.L. (1994). Statistics : Harcourt Brace College Publishers (Fifth edition)

On a effectué les calculs suivants :

	g0	g1	g2
Moy	6.2	12.0	15.4
Var	8.56	5.20	11.84

La moyenne générale des 15 scores est 11.2

A. Calculs (6 points)

1. La variance inter (*Vinter*) est égale à 14.43. Indiquer comment calculer cette valeur (rappeler la formule utilisée et développer les calculs) :

$$Vinter = \sum p_g \times (m_g - m)^2 = 0.33 \times (6.2 - 11.2)^2 + 0.33 \times (12.0 - 11.2)^2 + 0.33 \times (15.4 - 11.2)^2 = 14.43$$

$$Vinter = \sum p_g \times (m_g - m)^2 = \frac{(6.2 - 11.2)^2 + (12.0 - 11.2)^2 + (15.4 - 11.2)^2}{3}$$

2. La variance intra (*Vintra*) est égale à 8.53. Indiquer comment calculer cette valeur (rappeler la formule utilisée et développer les calculs) :

$$Vintra = \sum p_g \times Var_g = (0.33 \times 8.56) + (0.33 \times 5.20) + (0.33 \times 11.84) = \frac{8.56 + 5.20 + 11.84}{3} = 8.53$$

3. Le rapport de corrélation, *Eta²*, est égal à 0.63. Indiquer comment calculer cette valeur (rappeler la formule utilisée et développer le calcul) :

$$Eta^2 = \frac{Vinter}{Vtotale} = \frac{14.43}{14.43 + 8.53} = \frac{14.43}{22.96}$$

B. Rapport de corrélation (6 points)

1. Rappeler ce que mesure *Vinter* :

Vinter *mesure la dispersion des moyennes des groupes*

2. Rappeler ce que mesure *Vintra* :

Vintra *mesure la dispersion intra-groupes (la dispersion des individus à l'intérieur de leur groupe)*

3. Imaginer un ou plusieurs changements dans les données du groupe g0, qui conduiraient à une *Vintra* plus grande sans modifier *Vinter*.

	Valeurs observées	Nouvelles valeurs
e01	4	4
e02	8	8
e03	**11**	**12**
e04	5	5
e05	**3**	**2**

Expliquer le raisonnement qui vous a conduit à proposer ce(s) changement(s) :

En diminuant la plus petite valeur (3) de 1 point et en augmentant la plus grande valeur (11) de 1 point également, par exemple, on augmente la dispersion à l'intérieur du groupe, donc on augmente la variance du groupe, donc on augmente Vintra.
En modifiant le minimum et le maximum de la même quantité (ici 1 point) on ne change pas la somme du groupe, donc on ne change pas sa moyenne, donc on ne change pas Vinter.

4. Imaginer un ou plusieurs changements dans les données du groupe g0 qui conduiraient à une *Vinter* plus grande sans changer la valeur de *Vintra*.

	Valeurs observées	Nouvelles valeurs
e01	4	3
e02	8	7
e03	11	10
e04	5	4
e05	3	2

Expliquer le raisonnement qui vous a conduit à proposer ce(s) changement(s) :

En diminuant toutes les valeurs du groupe on baisse de la moyenne de ce groupe qui avait déjà la plus faible moyenne. En faisant cela on éloigne un peu plus la moyenne de ce groupe des autres. On augmente donc Vinter.
Par ailleurs, le fait de diminuer toutes les valeurs du groupe de la même quantité (ici 1 point) ne modifie pas la variance de ce groupe (cf. propriétés affines de la variance) donc ne modifie pas Vintra.

C. Inférence (8 points)

1. Définir précisément la population parente :

Il s'agit des enfants de 7 ans de cette ville anglaise.

2. Indiquer en quoi consiste l'hypothèse nulle (H0) classique (testée par le test *F*) :

H0 : Dans la population parente, la maturité sociale des enfants de 7 ans est la même quel que soit le nombre d'années de préscolarisation effectuées.

3. Le test *F* de Fisher-Snédécor donne les résultats suivants :
ddl1 = 2 *ddl2* = 12 *F* = 10.14 *p* = 0.26%

a. Indiquer le résultat du test (Significatif / Non Significatif) en justifiant votre réponse :

Le test est significatif car le seuil observé p (0.26%) est inférieur à la valeur repère usuelle (5%).

b. Peut-on rejeter l'hypothèse nulle énoncée précédemment ?

Oui

c. Rédiger une conclusion inférentielle :

Il semble que, chez l'ensemble des enfants de 7 ans de cette école de cette ville anglaise, le niveau de maturité sociale est lié au nombre d'années de préscolarisation (F[2 ;12]=10.14, p=0.26% < 5%, Significatif) .

Utilisation des tables (4 points)

On a reproduit ci-dessous une partie d'une table des distributions du *T* de Student.

p*/2** ***p **ddl**	**.025 (2.5%)** **.05 (5%)**	**.005 (0.5%)** **.01 (1%)**	**.0005 (0.05%)** **.001 (0.1%)**
1	12.71	63.66	636.62
2	4.30	9.92	31.60
3	3.18	5.84	12.92
4	2.78	4.60	8.61
5	2.57	4.03	6.87

Indiquer, pour chacun des tests suivants (*t* et degrés de liberté), une valeur approchée du seuil *p* (par exemple 1% < *p* < 5%) et le résultat du test (Significatif / Non significatif).

t	*Degrés de liberté*	Seuil *p* approché	Résultat du test
12.00	1	*p* > 5%	NS
60.00	1	1% < p < 5%	S
60.00	2	p < 0.1%	S
32.00	2	p < 0.1%	S
5.00	3	1% < p < 5%	S
2.00	4	p > 5%	NS

Phi² et *Khi²* (4 points)

1. A quel type de tableau s'appliquent ces deux statistiques ? Donner deux synonymes de ce type de tableau :

Tableau de contingence, Tri croisé, Tableau de correspondance, Contingency table, Distribution bivariée, Cross tabulated data, Tableau de dépendance, Table Banner.

2. Indiquer deux aspects sur lesquels *Phi²* et *Khi²* se différencient radicalement :

a. Phi² est une statistique descriptive alors que Khi² est une statistique inférentielle.

b. Phi² mesure l'ampleur de la liaison (observée) alors que Khi² évalue seulement l'existence de la liaison (parente)

Dossier Lecture (12 points)

Source : données fournies par A.N. Menchikoff.

Huit enfants sélectionnés dans une classe de CM1 sont invités à lire un texte à haute voix, à leur propre rythme. Après la lecture, ils subissent une épreuve de compréhension. On relève le temps (TEMPS) mis en secondes pour lire ce texte, puis on compte le nombre de bonnes réponses (BR) données à cette épreuve de compréhension. Les résultats sont présentés dans le tableau et le graphique ci-après.

Question initiale : on aimerait savoir d'une part, s'il existe une relation entre le temps de lecture du texte et sa compréhension par le lecteur, et d'autre part si on peut prédire le nombre de BR à l'épreuve de compréhension à partir du temps de lecture.

	e1	e2	e3	e4	e5	e6	e7	e8
TEMPS	57	53	51	60	65	71	75	82
BR	7	13	14	12	8	5	8	4

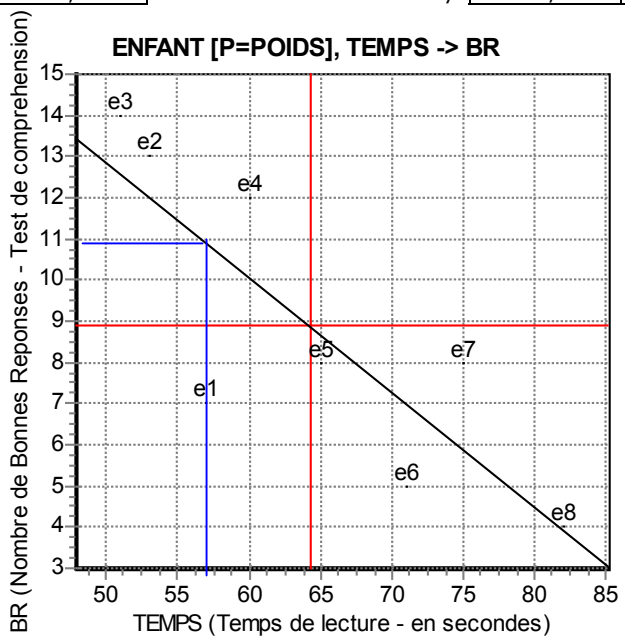
On donne les résultats suivants :

Moyennes

	TEMPS	BR
Moy	64,3	8,9

Écarts-type

	TEMPS	BR
Ety	10,3	3,5



Corrélation (7 points)

1. Tracer les axes moyens sur le graphique précédent.

2. Pour l'enfant e1, ses deux scores Z (pour les deux variables TEMPS et BR) sont :
TEMPSZ = -0.70 et BRZ = -0.54
a. Indiquer comment ont été calculées ces 2 valeurs :

$$TEMPSZ(e1) = \frac{x_{e1} - Moy_{TEMPS}}{Ety_{TEMPS}} = \frac{57 - 64.3}{10.3}$$

$$BRZ(e1) = \frac{x_{e1} - Moy_{BR}}{Ety_{BR}} = \frac{7 - 8.9}{3.5}$$

b. Quel est, pour cet enfant e1, le signe de sa contribution à la covariance et à la corrélation (justifier) :

Sa contribution à la covariance est positive. En effet :
- e1 a deux scores (TEMPS et BR) situés en dessous de la moyenne du groupe,
- ses écarts à la moyenne (et ses scores Z) sont donc négatifs,
- la contribution à la covariance a le signe du produit des écarts à la moyenne (= le signe du produit des scores Z),
- or ce produit est positif (- × - = +).

c. En quoi l'examen du graphique pouvait laisser prévoir ce résultat ?

L'individu e1 est situé dans le quadrant en bas à gauche, ce qui signifie qu'il est en dessous de la moyenne pour les deux scores.

3. Sachant que le coefficient de corrélation de Bravais-Pearson est égal à –0.83 :

a. Quelle serait la valeur de ce coefficient si les scores étaient divisés par 2 ?

Rbp = -0.83 (le coefficient de corrélation est invariant (au signe près) pour toute transformation affine (de type Y = a.X + b) d'une ou des deux variables).

b. Quelle serait la valeur de ce coefficient si les temps étaient multipliés par –1 ?

Rbp = +0.83 (le signe de la corrélation serait inversé)

Régression (5 points)

L'équation de la droite de régression est la suivante : ***BRpred = -0,28 x TEMPS + 26.8***
Cette droite de régression a été représentée sur le graphique de corrélation.

1. Indiquer pourquoi, compte tenu des résultats précédents, on devait s'attendre à obtenir un coefficient de régression (–0,28) négatif :

Le coefficient de régression a toujours le même signe que la corrélation linéaire de Bravais-Pearson (ici -0.83).
C'est également le signe de la pente de la droite de régression (cf. graphique).

2. Estimer graphiquement le nombre de bonnes réponses prédites (*BRpred*) pour l'enfant e1 par l'équation de régression, compte tenu de son temps de lecture. Indiquer, sur le graphe de corrélation, la procédure utilisée et reporter ci-dessous la valeur estimée graphiquement :

BRpred = 10.8 (entre 10.5 et 11.0)

3. En déduire une estimation de l'erreur de prédiction (*e*) :

e = 7 - 10.8 = -3.8 (entre -3.5 et -4.0)

4. Indiquer ci-dessous la procédure à utiliser (sans calculer) pour trouver précisément la prédiction *BRpred* pour le sujet e1 :

BRpred = (-0,28 × 57) + 26.8

Choix des statistiques (9 points)

Soient les 3 situations (3 types de données) suivantes où l'on doit analyser les données :
- Deux variables numériques
- Deux variables nominales
- Une variable nominale et une variable numérique.

Pour chacune de ces situations :

a. Citer 2 statistiques descriptives (on donnera uniquement leurs noms usuels complets, sans formule) qui évaluent l'ampleur du lien global entre les deux variables.

b. Citer une statistique inférentielle (on donnera uniquement son nom usuel complet, sans formule) qui permet de tester l'hypothèse nulle d'une absence de liaison entre les deux variables.

Situations	Deux statistiques descriptives	Une statistique inférentielle
Deux variables numériques	- Covariance - Coefficient de corrélation linéaire de Bravais-Pearson - Coefficient de détermination R^2	*T* de Student
Deux variables nominales	- *Phi²* - *V²* de Cramér	*Khi²*
Une variable nominale (G > 2) et une variable numérique	- *ECG* - *Eta²* - *Vinter, S²inter…*	*F* de Fisher-Snédécor

Intervalles de confiance (4 points)

Lors d'une analyse de données où il étudie les liaisons entre des variables numériques prises 2 à 2, un chercheur calcule, pour 4 couples de variables :
- un test T de Student pour tester l'hypothèse nulle d'une absence de liaison linéaire entre les deux variables,
- un intervalle de confiance sur la corrélation parente.

On rapporte ici uniquement les intervalles de confiance obtenus. En déduire quelle a été, pour chaque couple de variables, la conclusion du chercheur sur les 3 points suivants :
- résultat du test,
- signe de la corrélation parente,
- importance de la corrélation parente.

Intervalle de confiance	Résultat du test Significatif / Non significatif	Corrélation parente Négative/ Positive / Incertitude	Corrélation parente Faible / Modérée / Forte / Incertitude
[-.12 ; +.18]	*NS*	*Incertitude*	*Faible*
[-.35 ; +.25]	*NS*	*Incertitude*	*Faible ou Modérée (Incertitude)*
[-.75 ; -.45]	*S*	*Négative*	*Forte*
[+.25 ; +.55]	*S*	*Positive*	*Modérée ou forte (Incertitude)*

Combinaisons (3 points)

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 $n = 3$ pouvant être constitués à partir d'une population de taille $N = 7$ (indiquer le détail des calculs) :

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{7!}{3! \times 4!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (4 \times 3 \times 2 \times 1)} = \frac{7 \times 6 \times 5}{3 \times 2 \times 1} = \frac{7 \times 5}{1} = 35$$

Variable centrée-réduite ou score *Z* (4 points)

1. Rappeler la formule classique qui permet de calculer un score *Z* à partir de *X*

$$Z = \frac{X - moy}{ety}$$

2. Montrer que Z peut s'écrire sous une autre forme qui met en évidence qu'il s'agit d'une transformation affine (de type $a\ X + b$). Pour cela, indiquer les (2 ou 3) étapes de transformation de la formule précédente :

$$Z = \frac{X - moy}{ety} = \frac{X}{ety} - \frac{moy}{ety} = \frac{1}{ety} \times X - \frac{moy}{ety}$$

3. Indiquer alors, pour conclure, quelles sont les formules des coefficients *a* et *b* ainsi mis en évidence :

$$a = \frac{1}{ety}$$

$$b = -\frac{moy}{ety}$$