Université Paris Descartes (Paris 5) – Licence de Psychologie - 2008/2009
UE ADP2 – Responsable : Denis Corroyer - Première session : 7 janvier 2009

Durée de l'épreuve : 1h30.
Aucun document n'est autorisé. La calculette n'est pas autorisée.
Les différents exercices (encadrés) sont indépendants.
Le barème, indicatif, est sur 60 ; la note finale sera donnée sur 20.
Indiquer les réponses exclusivement sur ce document.
Ne rien écrire dans les marges.

CORRIGÉ

Les calculs et les graphiques ont été réalisés avec le logiciel SES-Pegase.

1. Dossier NATAL (14 points)

Le tableau suivant contient la liste des 14 pays d'Amérique du Nord et d'Amérique Centrale, dont la population dépassait le million d'habitants en 1985. Pour chaque pays on a :

- le taux d'urbanisation, noté URBAN (pourcentage de la population vivant dans des villes de plus de 100 000 habitants)

- le taux de natalité, noté NATAL (nombre de naissances par année pour 1 000 personnes).

Les deux premières colonnes correspondent aux données recueillies. Les deux colonnes suivantes correspondent à des valeurs calculées à partir de ces données.

PAYS	URBAN	NATAL	$NATAL_{pred}$	$NATAL_{res}$
CANADA	55.0	16.2	21.1	-4.85
COSTA RICA	27.3	30.5	32.1	-1.60
CUBA	33.3	16.9	29.7	-12.8
USA	56.5	16.0	20.5	-4.45
ELSALVADOR	11.5	40.2	38.4	1.80
GUATAMALA	14.2	38.4	37.3	1.07
HAITI	13.9	41.3	37.4	3.85
HONDURAS	19.0	43.9	35.4	8.49
JAMAIQUE	33.1	28.3	29.8	-1.49
MEXIQUE	43.2	33.9	25.8	8.14
NICARAGUA	28.5	44.2	31.6	12.6
TRINIDADE	6.80	24.6	40.3	-15.7
PANAMA	37.7	28.0	28.0	0.05
REPDOMINIC	37.1	33.1	28.2	4.91

On fait l'hypothèse que le taux de natalité d'un pays dépend, au moins en partie, de son taux d'urbanisation.

Source : Birkes & Dodge (1993) – Alternative Methods of Regression, New York, John Wiley and Sons, d'après Dodge (1999) - Analyse de régression appliquée, Paris : Dunod.Lecture du graphe

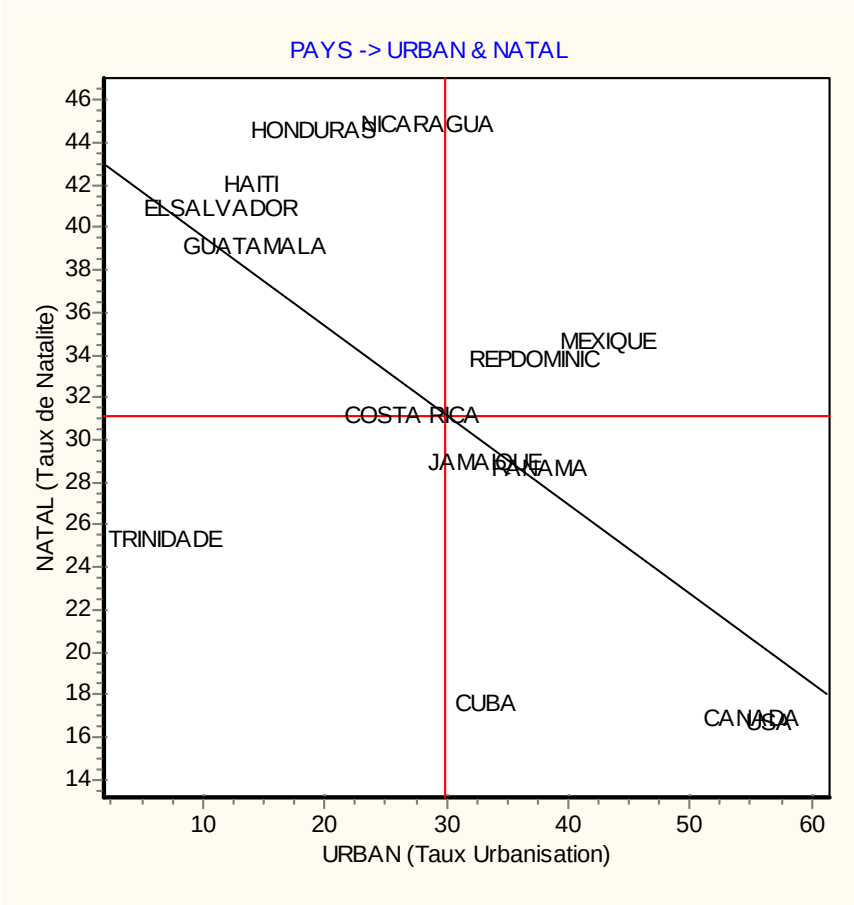
Lecture du graphe de corrélation

1. En quoi le graphe ci-après permet de qualifier la liaison, entre les deux variables, de liaison négative ?

Le nuage tend à avoir, globalement, une pente négative : plus le taux d'urbanisation est élevé, plus le taux de natalité tend à être faible.
Les pays se trouvent principalement dans le quadrant en haut à gauche (faible taux d'urbanisation et taux élevé de natalité) et dans le quadrant en bas à droite (taux d'urbanisation élevé et taux de natalité faible).

2. Indiquer un pays (plusieurs réponses possibles) qui va à l'encontre de cette interprétation globale du graphe et décrire brièvement la <u>particularité de ce pays</u> :

Certains pays sont caractéristiques d'une liaison positive :
- TRINIDADE a, par exemple, a la fois un <u>taux d'urbanisation faible</u> (en dessous de la moyenne du groupe) <u>ET</u> un <u>taux de natalité faible</u> (en dessous de la moyenne du groupe)
- Le MEXIQUE (et la république dominicaine) a, par exemple, à la fois un <u>taux d'urbanisation élevé</u> (au dessus de la moyenne du groupe) <u>ET</u> un <u>taux de natalité élevé</u> (au dessus de la moyenne du groupe)



3. L'intersection des deux axes indique le centre de gravité du nuage. En déduire :

a/ une <u>approximation du taux d'Urbanisation moyen</u> de ces 14 pays :

Moy (URBAN) ≈ 30 pour 100 000

b/ une <u>approximation du taux de Natalité moyen</u> de ces 14 pays :

Moy (NATAL) ≈ 31 pour 1 000

4. Que signifie le fait que <u>deux pays, comme le CANADA et les USA, apparaissent superposés</u> sur le graphe ?

Cela signifie qu'ils ont les mêmes coordonnées sur les deux variables,
c'est à dire qu'ils ont des <u>taux d'urbanisation et de natalité</u> proches.
C'est le cas de USA et CANADA.

Droite de régression

L'équation de régression linéaire est la suivante : $NATAL_{pred}$ = -0.40 x URBAN + 43

1. Que nous apprend le coefficient 43 ?

Ce coefficient prédit un taux de natalité de 43 naissances pour 1000 personnes si le taux d'urbanisation est nul (URBAN = 0 soit aucune ville de plus de 100 000 habitants)

2. Compte tenu de la valeur de ce coefficient et des propriétés générales d'une droite de régression, tracer la droite de régression sur le graphe précédent (sans calcul supplémentaire).

Remarque : la droite de régression passe par l'ordonnée à l'origine (de coordonnée (0, 43)) et par le centre de gravité du nuage (intersection des axes moyens).

Prédiction

On s'intéresse au cas particulier du CANADA (on rappelle que $NATAL_{pred}$ = -0.40 x URBAN + 43).

1. Indiquer comment ont été calculées les valeurs 21.1 et -4.85 (cf. tableau de présentation du dossier) pour le CANADA :

21.1 = -0.40 x 55.0 + 43
(remarque : 21.1 a été calculé précisément, avec toutes les décimales et non pas avec les valeurs arrondies présentées ici).)

-4.85 = 16.2 – 21.1 (même remarque)

2. Indiquer, en une phrase brève, à quoi correspond la valeur 21.1 :

21.1 est…
Le taux de Natalité estimé (prédit) pour le CANADA, à partir de son taux d'Urbanisation

3. Indiquer, en une phrase brève, à quoi correspond la valeur -4.85 :

-4.85 est…
l'écart entre le taux de Natalité réel et le taux de Natalité estimé (prédit) pour le CANADA (compte tenu de son taux d'Urbanisation).

Qualité de la prédiction

1. Pour cet exemple, comment pourrait-on calculer la valeur du coefficient R^2 (indiquer la formule générale, puis la formule adaptée à cet exemple) ?

R^2 = (Var (Y_{est}) / Var (Y) = Var ($NATAL_{est}$) / Var (NATAL)

2. On trouve R^2 =0.39. Interpréter, brièvement, cette valeur :

La connaissance du taux d'Urbanisation des pays permet d'estimer 39 % de la variance du Taux de Natalité de ces pays. Ce pourcentage peut être considéré comme important (R^2 > 16 %).

Inférence

Cela n'aurait guère de sens de mettre en œuvre des procédures inférentielles sur les données de ce dossier. Pourquoi ? (si besoin, on relira attentivement la présentation du dossier).

Le tableau de données porte sur la population parente ("les 14 pays d'Amérique du Nord et d'Amérique Centrale, dont la population dépassait le million d'habitants en 1985").
Quand bien même on considérerait que la population parente est constituée de l'ensemble des pays (de tous les continents) dont la population dépassait le million d'habitants, ce groupe de 14 pays ne constituerait pas un échantillon aléatoire de cette population parente et donc les procédures inférentielles ne pourraient être mises en oeuvre.

2. *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)

3. Choix des statistiques (9 points)

Soient les 3 situations (3 types de données) suivantes où l'on doit analyser les données :
- Analyser la liaison entre deux variables numériques
- Analyser la liaison entre deux variables nominales
- Analyser la liaison entre une variable nominale et une variable numérique.

Pour chacune de ces situations :

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

2. 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* *(Khi² de McNémar ou Khi² d'indépendance)*
Une variable nominale (G > 2) et une variable numérique	- *ECG* - *Eta²* - *Vinter, S²inter…*	*F* de Fisher-Snédécor

4. Test *t* de Student pour comparer 2 groupes indépendants (2 points)

1. Indiquer deux conditions préalables pour la mise en œuvre du test *T* de Student classique permettant de comparer les moyennes des deux groupes indépendants :

- Normalité des distributions parentes
- Tirage au hasard des unités dans la population
- Homogénéité des variances parentes

5. Intervalles de confiance (3 points)

Lors de recherches où il a comparé les QI de deux groupes indépendants, un chercheur a calculé :
- un test T de Student pour tester l'hypothèse nulle d'une absence de différence entre les deux moyennes parentes,
- un intervalle de confiance sur la différence parente.
On rapporte ici uniquement les intervalles de confiance obtenus pour la différence parente.

Il considère qu'une différence inférieure à 5 points de QI est faible, et qu'une différence supérieure à 10 points de QI est importante.
En déduire quelle a été, pour chaque couple de variables, la conclusion du chercheur sur les 3 points suivants :
- résultat du test (Significatif / Non significatif)
- signe de la différence parente (Négatif / Positif / Incertitude)
- importance de la différence parente (Faible / Forte / Incertitude).

Intervalle de confiance	Résultat du test : Significatif/Non significatif	Signe de la différence parente : Négatif / Positif / Incertitude	Importance de la différence parente : Faible / Forte / Incertitude
[-12 ; +18]	*NS*	*Incertitude*	*Incertitude*
[+12 ; +19]	*S*	*Positif*	*Forte*
[-4 ; +2]	*NS*	*Incertitude*	*Faible*
[+9 ; +15]	*S*	*Positif*	*Incertitude*

6. Mesures répétées sur une variable binaire (4 points)

Lors d'une expérience portant sur les effets d'un apprentissage sur la réussite à une tâche, on obtient le tableau suivant qui indique la répartition (en effectifs) des sujets selon leur réussite avant et après l'apprentissage :

		Après		
		Réussite	Échec	
Avant	Réussite	12	8	**20**
	Échec	18	12	**30**
		30	**20**	**50**

1. Calculer l'écart brut, en points de pourcentage, qui mesure l'évolution des réussites après l'apprentissage (indiquer le détail des calculs) :

Avant apprentissage : 20/50 = 40%
Après apprentissage : 30/50 = 60%

d = 60% -40% = 20 points de pourcentage

2. La formule du test *Khi²* de McNémar est la suivante : $Khi^2 = \frac{(|B-C|-1)^2}{B+C}$

a/ A quoi correspondent, les intitulés B et C dans cette formule ?

Ils correspondent aux « cases avec changement » c'est à dire aux effectifs de ceux qui ont progressé et de ceux qui ont régressé.

b/ Calculer la valeur du *Khi²* de McNémar pour ces données :

$$Khi^2 = \frac{(|B-C|-1)^2}{B+C} = \frac{(|18-8|-1)^2}{8+18} = \frac{9^2}{26} = 3.11$$

c/ Indiquer le nombre de degrés de liberté de ce test :

ddl = 1

7. Interprétation d'un test non significatif (2 points)

Un chercheur teste l'hypothèse qu'il n'existe PAS de différence, dans la population parente, entre les garçons et les filles sur les performances à une certaine tâche. Après avoir constaté l'existence d'une différence non nulle sur un échantillon, il met en œuvre un test statistique (*t* de Student ou *F* de Fisher-Snedecor) et constate que ce test est non significatif.
Quelle conclusion peut-il tirer de ce résultat ?

Ce résultat ne lui permet pas de confirmer son hypothèse.
Un test non significatif ne permet pas de se prononcer sur l'existence, ou l'absence, d'une différence dans la population.

Remarques :
Il n'existe en fait aucune procédure statistique permettant de confirmer l'absence de différence entre deux groupes. Tout au mieux peut-on tester l'hypothèse d'une différence faible.
Le chercheur devrait, par exemple, calculer un intervalle de confiance. Si cet intervalle ne comprend que des valeurs faibles, il pourrait alors au moins conclure que la différence est faible.

8. Variance intra et variance inter (5 points)

On trouve ci-dessous les moyennes et variances d'une variable numérique recueillie sur deux groupes d'effectifs différents (6 et 3) :

	Effectifs	Moyennes	Variances
g1	6	9	4
g2	3	12	7

1. Calculer la moyenne générale des 9 sujets (indiquer le détail des calculs) :

$$Moy = \frac{6\times9+3\times12}{9} = \frac{54+36}{9} = \frac{90}{9} = 10.00$$

2. Calculer la variance intra-groupes (indiquer le détail des calculs) :

$$Vintra = \frac{(6\times4)+(3\times7)}{9} = \frac{24+21}{9} = \frac{45}{9} = 5.00$$

3. Calculer la variance inter-groupes (indiquer le détail des calculs) :

$$Vinter = \frac{6\times(9-10)^2+3\times(12-10)^2}{9} = \frac{6+12}{9} = 2.00$$

9. Rédaction d'une conclusion (12 points)

Lors d'une expérience pédagogique portant sur des enfants de 12 ans scolarisés dans une école pour enfants à haut-potentiel, on compare, sur deux groupes indépendants de 15 enfants tirés au hasard dans cette école, l'efficacité de deux pédagogies, une pédagogie « concrète » (p1) et une pédagogie « formelle » (p2). On note leur temps de résolution (en minutes) d'une épreuve.

On considère qu'une différence de performance est :
- faible, si elle est inférieure à 2 minutes,
- importante, si elle est supérieure à 5 minutes.

On fait l'hypothèse que la pédagogie formelle est plus efficace.

Analyse descriptive (6 points)

On trouve Moy_{p1} = 34 mn et Moy_{p2} = 22 mn.
Compte tenu de ces éléments, rédiger une conclusion descriptive détaillée :

Sur ces 30 enfants de 12 ans scolarisés dans une école pour enfants à haut-potentiel,

on observe que leur temps de résolution d'une épreuve,

est meilleur après une pédagogie formelle (Moy = 22 mn)
par rapport à une pédagogie concrète (Moy=34 mn)

La différence entre les moyennes (12 minutes)

apparaît importante/forte/grande (> 5 minutes)

Analyse inférentielle (6 points)

Le test *t* de Student (et le test *F* de l'ANOVA) indique *p* = 0.6% et le calcul d'un intervalle de confiance sur la différence parente indique : IC(5%) = [8 ; 16].
Compte tenu de ces éléments, rédiger une conclusion inférentielle détaillée :

Il semble que l'on puisse conclure que,

chez l'ensemble des enfants de 12 ans de cette école pour enfants à haut-potentiel

il existe bien une différence de temps de résolution de l'épreuve

en faveur des enfants qui ont reçu une pédagogie formelle, par rapport à ceux qui ont reçu une pédagogie concrète.

(test Significatif, p= 0.6% < 5%)

De plus on peut conclure à une différence importante (supérieure à 5 minutes)

(IC(5%) = [8 ; 16])

10. Théorème de Bayes (5 points)

On indique ci-dessous la répartition en pourcentages, par groupe, des réponses à une question :

	r1	r2	r3	
G	30 %	50 %	20 %	100 %
F	60 %	10 %	30 %	100 %

Soit :
- *J* la variable Sexe (G/F)
- *K* la variable Réponse (r1/r2/r3)

1. On s'intéresse au pourcentage, indiqué dans le tableau, de *P*(r1/G).

a/ « Traduire » en langage naturel, cette expression :

P(r1/G) *: Pourcentage de G qui donnent la réponse r1*

b/ Rapporter ci-dessous ce pourcentage, indiqué dans le tableau :

P(r1/G) = 30%

2. On utilise le théorème de Bayes pour calculer P(G/r1) connaissant P(r1/G) :

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

a/ Recopier ci-dessous la formule en remplaçant *k* par sa valeur :

$$P(\text{G/r1}) = \frac{P(j)\cdot P(r1/j)}{\sum_J P(j)\cdot P(r1/j)}$$

b/ Développer la somme du dénominateur pour les deux modalités de *J* (*j*=G et *j*=F) :

$$P(\text{G/r1}) = \frac{P(j)\cdot P(r1/j)}{\sum_J P(j)\cdot P(r1/j)} = \frac{P(j)\cdot P(r1/j)}{P(G)\cdot P(r1/G)+P(F)\cdot P(r1/F)}$$

c/ Calculer *P*(G/r1) en faisant l'hypothèse que *P*(G) = 0.50 (50%) et *P*(F) = 0.50 (50%). Indiquer le détail des calculs :

$$P(\text{G/r1}) = \frac{P(G)\cdot P(r1/G)}{P(G)\cdot P(r1/G)+P(F)\cdot P(r1/F)} = \frac{0.50\times 0.30}{0.50\times 0.30+0.50\times 0.60}$$

$$= \frac{0.15}{0.15+0.30} = \frac{0.15}{0.45} = \frac{1}{3} = 0.33\,(33\%)$$

P(G/r1) = 33%