Code Module : MAT4055	Intitulé du Module : Analyse de données
Date : 23 mai 2008	Durée : 1 heure 30
Professeur : Maumy-Bertrand	Nbre de pages : 7
Evaluation: ☒	Test: ☐ Classe : 4B21-4B22
Documents autorisés :	Oui : ☒ Non : ☐
Calculatrice autorisée :	Oui : ☒ Non : ☐
Précision sur le barème si QCM :	
Commentaires : Vous devez indiquer sur la copie le sujet traité et le rendre.	

Nom & Prénom de l'étudiant :
Code étudiant : Classe :

SUJET :

Sujet numéro 1

Le sujet comporte trois exercices indépendants. Il est demandé à l'étudiant de ne traiter que deux exercices parmi les trois. La rédaction de trois exercices entraîne automiquement un rejet de correction. Les calculatrices sont autorisées ainsi que les notes de cours, les tds et leurs corrigés. Les tables de Fisher, de Student et du χ^2 doivent être distribuées.

Exercice 1. Le composant électronique.
Un certain composant électronique est fabriqué une fois par mois par l'entreprise Micro-Systèmes. La quantité fabriquée varie avec la demande du marché. Dans le but de planifier la production et d'établir certaines normes sur le nombre d'hommes-minutes exigés pour la production de différents lots de ce composant électronique, le responsable de la productoin a relevé l'information suivante pour 15 cédules de production. Le nombre d'hommes-minutes est identifié par Y et la quantité fabriquée par X.

Y_i	150	192	264	371	300	358	192	134	242	238	226	302	340	182	169
X_i	35	42	64	88	70	85	40	30	55	60	51	72	80	44	39

1. Le responsable de la production envisage d'utiliser le modèle linéaire simple comme modèle prévisionnel. Spécifier ce modèle et bien identifier chacune des composantes du modèle dans le contexte de ce problème.
2. Déterminer l'équation de la droite de régression.
3. Donner une estimation de $s(\beta_1)$.
4. Tester, au seuil de 5%, l'hypothèse nulle $\mathcal{H}_0$ suivante avec un test approprié (vous donnerez le nom de ce test) :
$$\mathcal{H}_0 : \beta_1 = 0$$
contre
$$\mathcal{H}_1 : \beta_1 \neq 0.$$
5. Donner la valeur de la variation qui est expliquée par la droite de régression et la variation qui est inexpliquée par la même droite.
6. Donner le pourcentage de variation qui est expliqué par la droite de régression.
7. Calculer une estimation du nombre moyen d'hommes-minutes requis pour : $X_h = 42; X_h = 57; X_h = 72$.
8. Pour quelle quantité X_n, l'estimation du nombre moyen d'hommes-minutes requis serait-elle la plus précise parmi les trois valeurs précédentes ?
9. Entre quelles valeurs peut se situer le vrai nombre moyen d'hommes-minutes requis pour les lots dont la quantité a été déterminée à la question précédente ? Utiliser un niveau de confiance de 95%.
10. Quelle est la marge d'erreur dans l'estimation effectuée à la question précédente ?

```
> quantite<-c(35,42,64,88,70,85,40,30,55,60,51,72,80,44,39)
> quantite
35 42 64 88 70 85 40 30 55 60 51 72 80 44 39
> nombre-hommes<-c(150,192,264,371,300,358,192,134,242,238,226,302,
340,182,169)
> nombre-hommes
150 192 264 371 300 358 192 134 242 238 226 302 340 182 169
> model<-lm(nombre-hommes∼quantite)
> residus<-residuals(model)
> shapiro.test(residus)
Shapiro-Wilk normality test
data : residus
W = 0.9809, p-value = 0.975
> summary(model)
Call : lm(formula = nombre-hommes ∼ quantite)
Residuals :
Min 1Q Median 3Q Max
-18.050 -4.167 1.534 4.908 16.283
Coefficients :
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.0511	7.3458	2.049	0.0612
quantite	4.0166	0.1227	32.723	$7.12e-14$

```
Residual standard error : 8.667 on 13 degrees of freedom
Multiple R-Squared : 0.988, Adjusted R-squared : 0.9871
F-statistic : 1071 on 1 and 13 DF, p-value : 7.119e-14
> anova(model)
Analysis of Variance Table
Response : nombre-hommes
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
quantite	1	80441	80441	1070.8	$7.119e-14$
Residuals	13	977	75		

Exercice 2. La vigne se traite.
Nous nous proposons de comparer l'efficacité de deux traitements T_1 et T_2 destinés à combattre une certaine maladie de la vigne. Dans un vignoble atteint de cette maladie, nous choisissons au hasard deux échantillons, l'un de 110 pieds de vigne l'autre de 90 pieds de vigne, auquels nous appliquons respectivement les traitements T_1 et T_2. Quelques mois après la fin des traitements, nous observons les résultats obtenus. À cet effet, nous partageons chacun des échantillons obtenus en trois catégories :

a) A : disparition totale de la maladie

b) B : présence de quelques séquelles

c) C : persistence de la maladie.

Les résultats obtenus figurent dans le tableau suivant :

	A	B	C
T_1	80	25	5
T_2	60	18	12

Les effets des deux traitements sont-ils significativement différents ?
Pour répondre à la question, vous effectuerez un test dont vous donnerez le nom, puis vous énoncerez les deux hyptohèses associées à ce test ainsi que la valeur de la statistique de ce test. Enfin, il manque deux valeurs dans la sortie de `R`, retrouvez cette valeur.

```
> vigne<-matrix(c(80,25,5,60,18,12),byrow=T,nrow=2,
dimnames=list(c("T1","T2"),c("A","B","C")))
> vigne
```

```
    A  B  C
T1 80 25  5
T2 60 18 12
```

```
> chisq.test(vigne,correct=FALSE)
Pearson's Chi-squared test
data : vigne
X-squared = 4.9283, df = 2, p-value = 0.08508
> chisq.test(vigne,correct=FALSE)$expected
```

```
    A     B    C
T1  ? 23.65 9.35
T2  ? 19.35 7.65
```

Exercice 3. Cultures pour BCG.
Disposant de 5 milieux A, B, C, D, E pour la culture du BCG, nous nous proposons de savoir si, dans l'ensemble, les milieux sont équivalents ou, au contraire, certains favorisent la croissance plus que d'autres (nombre colonies plus élevé).
Nous avons donc ensemencé, à partir d'une même suspension de BCG, 8 tubes par milieu de culture. Le tableau suivant donne le nombre de colonies obtenues pour chaque tube :

Milieu de culture Résultats	A	B	C	D	E
x_{i1}	10	11	7	12	7
x_{i2}	12	18	14	9	6
x_{i3}	6	13	9	7	7
x_{i4}	13	8	10	8	5
x_{i5}	9	15	9	13	6
x_{i6}	10	16	11	14	7
x_{i7}	8	9	7	10	9
x_{i8}	9	13	9	11	6

1. Le nombre de colonies peut-il être considéré comme distribué normalement dans chacune des populations de mesure ?
2. La variance du nombre de colonies dans chacune des populations de mesures peut-elle être considérée comme indépendante du milieu de culture ?
3. L'homogénéité des résultats est-elle vérifiée ? Si oui, comment procédez-vous pour répondre à cette question ?
4. Les conditions d'application du modèle linéaire sont-elles vérifiées ? Si oui, expliquer votre réponse.
5. Donner le tableau de l'analyse de la variance.
6. Quelle est la valeur estimée de la variance résiduelle ?
7. Au risque de 5%, pouvons-nous considérer que le milieu de culture possède une une influence sur la croissance du BCG ?
8. Pouvons-nous séparer les milieux de culture en groupes ne présentant pas de différence significative au seuil de 5% ? Si oui, expliquer comment vous procédez.
9. Dans le cas où vous avez répondu dans l'affirmative à la question précédente, faire cette répartition en groupes homogènes, en indiquant les milieux et les moyennes correspondantes du nombre de colonies.

```
> options(contrasts=c("contr.sum","contr.poly"))
> milieu<-rep(1 :5,c(8,8,8,8,8))
> milieu
1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5
> resultats<-c(10,12,6,13,9,10,8,9,11,18,13,8,15,16,9,13,7,14,9,10,9,
11,7,9,12,9,7,8,13,14,10,11,7,6,7,5,6,7,9,6)
> resultats
10 12 6 13 9 10 8 9 11 18 13 8 15 16 9 13 7 14 9 10 9
11 7 9 12 9 7 8 13 14 10 11 7 6 7 5 6 7 9 6
> milieu<-factor(milieu)
> BCG<-data.frame(milieu,resultats)
> str(BCG)
'data.frame' : 40 obs. of 2 variables :
milieu : Factor w/ 5 levels "1","2","3","4",.. : 1 1 1 1 1 1 1 1 2 2 ...
resultats : num 10 12 6 13 9 10 8 9 11 18 ...
> rm(milieu)
> rm(resultats)
> moy<-tapply(BCG$resultat,BCG$milieu,mean)
> moy
```

1	2	3	4	5
9.625	12.875	9.500	10.500	6.625

```
> model1<-lm(resultats~milieu,BCG)
> model1
Call :
lm(formula = resultats ~ milieu, data = BCG)
Coefficients :
```

(Intercept)	milieu1	milieu2	milieu3	milieu4
9.825	−0.200	3.050	−0.325	0.675

```
> residus<-residuals(model1)
> shapiro.test(residus)
Shapiro-Wilk normality test
data : residus
W = 0.9795, p-value = 0.6711
> bartlett.test(residus~milieu,BCG)
Bartlett test of homogeneity of variances
data : residus by milieu
Bartlett's K-squared = 6.6713, df = 4, p-value = 0.1543
> summary(model1)
Call :
lm(formula = resultats ~ milieu, data = BCG)
Residuals :
```

Min	1Q	Median	3Q	Max
−4.8750	−1.5312	−0.1875	1.5000	5.1250

```
Coefficients :
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.8250	0.3823	25.699	$< 2e-16$
milieu1	−0.2000	0.7646	−0.262	0.795188
milieu2	−0.1250	0.7646	3.989	0.000322
milieu3	0.8750	0.7646	−0.425	0.673404
milieu4	−3.0000	0.7646	0.883	0.383372

```
Residual standard error : 2.418 on 35 degrees of freedom
Multiple R-Squared : 0.4406, Adjusted R-squared : 0.3766
F-statistic : 6.891 on 4 and 35 DF, p-value : 0.0003354
> anova(model1)
Analysis of Variance Table
Response : resultats
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
milieu	4	161.150	40.287	6.891	0.0003354
Residuals	35	204.625	5.846		

```
> model2<-aov(resultats milieu,BCG)
> TukeyHSD(model2)
Tukey multiple comparisons of means
95% family-wise confidence level
Fit : aov(formula = resultats   milieu, data = BCG)
milieu
```

	diff	lwr	upr	p adj
2 − 1	3.250	−0.2258602	6.7258602	0.0762105
3 − 1	−0.125	−3.6008602	3.3508602	0.9999728
4 − 1	0.875	−2.6008602	4.3508602	0.9495568
5 − 1	−3.000	−6.4758602	0.4758602	0.1181661
3 − 2	−3.375	−6.8508602	0.1008602	0.0605230
4 − 2	−2.375	−5.8508602	1.1008602	0.3040122
5 − 2	−6.250	−9.7258602	−2.7741398	0.0000896
4 − 3	1.000	−2.4758602	4.4758602	0.9203980
5 − 3	−2.875	−6.3508602	0.6008602	0.1453756
5 − 4	−3.875	−7.3508602	−0.3991398	0.0225643