Code Module : 4056 **Intitulé du Module : Analyse de données**

Date : mai 2015 **Durée : 1 heure 30**

Professeur : Mme Bertrand Myriam **Nombre de pages: 5**

Examen: **X** **Contrôle:** ☐ **Classe: 4ème année SI**

Documents autorisés : **Oui** ☒ **Non** ☐
Calculatrice autorisée : **Oui** ☒ **Non** ☐
Ordinateur autorisé : **Oui** ☐ **Non** ☒

Précision sur le barème si QCM :

Commentaires :

NOM de l'étudiant:

Prénom de l'étudiant:

Code étudiant :

Examen de Rattrapage

— Vous prendrez un soin particulier à préciser quelles sont les hypothèses testées.
— Tous les tests seront effectués au seuil de signification $\alpha = 5\%$.

Le sujet comporte trois exercices indépendants.

Exercice 1. Contrôle de qualité et campagne marketing associée.

Laisser tomber son smartphone peut être un drame : on perd ses contacts, tout lien au monde, et en plus, cet accident n'est pas couvert par la garantie et il faudra débourser plusieurs centaines d'euros pour un smartphone de remplacement. Bref, voilà un créneau tout trouvé pour Lemon, un concurrent d'Apple et d'Orange : cette entreprise a construit un smartphone révolutionnaire qui peut encaisser jusqu'à 260 chocs en moyenne, selon l'argumentaire commercial développé. Sceptique, une association civique et féminine d'usagers de téléphonie mobile, « Téléphonées », mène un test sur 81 appareils de Lemon, dont les résultats sont reproduits ci-dessous :

Échantillon	$n = 81$ téléphones
Moyenne des nombres de chocs avant panne	$\widehat{\mu} = 264,9$
Écart-type des nombres de chocs	$s_{n,c} = 24,4$

Que faut-il en conclure ?
Vous pouvez répondre à la question posée par exemple avec le test adéquat.

```
> qt(0.975,64)
[1] 1.99773
> qt(0.975,63)
[1] 1.998341
> qt(0.95,64)
[1] 1.669013
> qt(0.95,63)
[1] 1.669402
> qnorm(0.95)
[1] 1.644854
> qnorm(0.975)
[1] 1.959964
```

Exercice 2. Le Titanic.

Le naufrage du Titanic a été un grand drame humain. Mais certains s'en seraient-ils mieux sortis que d'autres ? Considérez les données suivantes (ce sont les vraies données !) :

Survécu	Non	Oui
Membres d'équipage	673	212
Première classe	122	203
Deuxième classe	167	118
Troisième classe	528	178

Que pensez-vous donc de cette affirmation « Mais certains s'en seraient-ils mieux sortis que d'autres ! » ? Partagez-vous cette idée ?
Pour répondre à la question posée, vous pouvez réaliser le test adéquat.

```
> tableau<-matrix(c(673,212,122,203,167,118,528,178),nrow=4,byrow=T)
> tableau
     [,1] [,2]
[1,]  673  212
[2,]  122  203
[3,]  167  118
[4,]  528  178
> chisq.test(tableau)

        Pearson's Chi-squared test

data:  tableau
X-squared = 190.4011, df = 3, p-value < 2.2e-16
> qchisq(0.975,2)
[1] 7.377759
> qchisq(0.95,2)
[1] 5.991465
> qchisq(0.975,3)
[1] 9.348404
> qchisq(0.95,3)
[1] 7.814728
> qchisq(0.975,6)
[1] 14.44938
> qchisq(0.95,6)
[1] 12.59159
> qchisq(0.975,8)
[1] 17.53455
> qchisq(0.95,8)
[1] 15.50731
```

Exercice 3. Le salaire dépend il du niveau d'études en formation initiale ?

Le tableau suivant présente les salaires annuels bruts d'individus au bout de cinq ans d'expériences selon leur niveau de formation initiale. Qu'en pensez-vous ?

Licence	Master/École d'ing.	Doctorat
35,9	39,7	25,6
32,5	32,6	48,2
36,0	25,7	47,3
28,1	35,4	29,3
22,4	29,1	35,6
23,5	40,3	26,4
24,6	27,6	28,6
21,5	22,1	47,5
24,2	28,9	35,8
23,7	31,6	42,6
30,7	32,5	45,0

1. Proposer un modèle statistique qui permet d'étudier une relation (préciser le type de relation) entre le salaire annuel brut et le niveau de formation initiale. Préciser la nature de chacune des variables présentes dans le modèle statistique proposé.
2. Les conditions d'application du modèle linéaire sont-elles vérifiées ? Si oui, expliquer votre réponse.
3. Donner le tableau de l'analyse de la variance.
4. D'après les sorties statistiques réalisées avec le logiciel R qui se trouvent ci-dessous, pouvez-vous conclure à une éventuelle significativité du niveau de formation initiale sur le salaire annuel brut ? Pour répondre à cette question, utiliser un test. Vous citerez le nom du test, les hypothèses, la statistique du test et donnerez la conclusion du test (vous préciserez quelle règle vous utilisez).
5. Pouvez-vous séparer les niveaux de formation initiale en groupes ne présentant pas de différence significative au seuil de 5% ? Si oui, expliquer comment vous procédez.

```
> niveauetude<-rep(c(1,2,3),c(11,11,11))
> niveauetude<-factor(niveauetude)
> salaire<-c(35.9,32.5,36,28.1,22.4,23.5,24.6,21.5,24.2,23.7,30.7,39.7,
32.6,25.7,35.4,29.1,40.3,27.6,22.1,28.9,31.6,32.5,25.6,48.2,47.3,29.3,
35.6,26.4,28.6,47.5,35.8,42.6,45)
> jeudedonnee<-data.frame(niveauetude,salaire)
> str(jeudedonnee)
'data.frame':   33 obs. of  2 variables:
 $ niveauetude: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ salaire    : num  35.9 32.5 36 28.1 22.4 23.5 24.6 21.5 24.2 23.7 ...
> rm(niveauetude)
> rm(salaire)
```

```
> modele1<-aov(salaire~niveauetude,data=jeudedonnee)
> modele1
Call:
   aov(formula = salaire ~ niveauetude, data = jeudedonnee)

Terms:
                niveauetude Residuals
Sum of Squares     546.7927 1412.4036
Deg. of Freedom           2        30

Residual standard error: 6.861496
Estimated effects may be unbalanced
> residus<-residuals(modele1)
> shapiro.test(residus)

        Shapiro-Wilk normality test

data:  residus
W = 0.9519, p-value = 0.1511
> bartlett.test(residus~niveauetude,data=jeudedonnee)

        Bartlett test of homogeneity of variances

data:  residus by niveauetude
Bartlett's K-squared = 3.3808, df = 2, p-value = 0.1844
> summary(modele1)
            Df Sum Sq Mean Sq F value  Pr(>F)
niveauetude  2  546.8  273.40   5.807 0.00738
Residuals   30 1412.4   47.08
> TukeyHSD(modele1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = salaire ~ niveauetude, data = jeudedonnee)

$niveauetude
        diff       lwr      upr     p adj
2-1 3.854545 -3.358226 11.06732 0.3967629
3-1 9.890909  2.678138 17.10368 0.0055781
3-2 6.036364 -1.176408 13.24913 0.1148140
```