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

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

Professeur : Mme Bertrand Myriam **Nombre de pages: 9**

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 d'analyse de données

Durée : 1 heure 30.

Le sujet comporte trois exercices indépendants.

— Les calculatrices sont autorisées.
— Le cours, les exercices de travaux dirigés, leurs corrigés ainsi que les notes de cours sont autorisés. Tout autre document est interdit
— Afin de pouvoir traiter les questions, plusieurs résultats numériques et graphiques ont été intégrés au document.
— Vous prendrez un soin particulier à préciser quelles sont les hypothèses testées.
— Tous les tests seront effectués au seuil de signification $\alpha = 5\%$.

Exercice 1. Traitement contre l'urée. Cet exercice comporte six questions.

Cinq centres hospitaliers utilisent un traitement différent pour combattre le taux élevé d'urée dans le sang chez les malades atteints de lésions rénales. Le caractère étudié est le taux d'urée (en décigrammes par litre de sang) après traitement. Dans chaque centre hospitalier, nous l'avons mesuré chez sept patients. Les données sont présentées ci-dessous.

traitement 1	traitement 2	traitement 3	traitement 4	traitement 5
4,5	7,5	8,0	2,0	6,5
2,5	3,0	6,5	7,5	5,5
6,0	2,5	6,0	4,0	6,0
4,5	4,0	3,5	2,5	4,5
3,0	2,0	5,0	5,0	4,0
5,5	4,0	7,0	3,5	7,0
3,5	5,5	5,0	6,5	5,5

1. Proposer un modèle statistique qui permet d'étudier une relation (préciser le type de relation) entre le taux d'urée dans le sang et le traitement. 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 traitement sur le taux d'urée dans le sang ? 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 traitements en groupes ne présentant pas de différence significative au seuil de 5% ? Si oui, expliquer comment vous procédez.

6. Dans le cas où vous avez répondu dans l'affirmative à la question précédente, faire cette répartition en groupes homogènes, en indiquant les traitements et les moyennes correspondantes du taux d'urée dans le sang.

```
> traitement<-rep(1:5,c(7,7,7,7,7))
> traitement<-factor(traitement)
> taux<-c(4.5,2.5,6,4.5,3,5.5,3.5,7.5,3,2.5,4,2,4,5.5,8,6.5,6,3.5,5,
7,5,2,7.5,4,2.5,5,3.5,6.5,6.5,5.5,6,4.5,4,7,5.5)
> exo1<-data.frame(traitement,taux)
> str(exo1)
'data.frame':   35 obs. of  2 variables:
 $ traitement: Factor w/ 5 levels "1","2","3","4",
 ..: 1 1 1 1 1 1 1 2 2 2 ...
 $ taux      : num  4.5 2.5 6 4.5 3 5.5 3.5 7.5 3 2.5 ...
> mean<-tapply(exo1$taux,exo1$traitement,mean)
> mean
       1        2        3        4        5
4.214286 4.071429 5.857143 4.428571 5.571429
> var<-tapply(exo1$taux,exo1$traitement,var)
> var
       1        2        3        4        5
1.654762 3.619048 2.226190 4.119048 1.119048
> boxplot(taux~traitement)
```



```
> modele1<-aov(taux~traitement,data=exo1)
> modele1
Call:
   aov(formula = taux ~ traitement, data = exo1)

Terms:
                traitement Residuals
Sum of Squares    19.04286  76.42857
Deg. of Freedom          4        30

Residual standard error: 1.596126
Estimated effects may be unbalanced

> options(contrasts=c("contr.sum","contr.poly"))
> modele2<-lm(taux~traitement,data=exo1)
> summary(modele3)

Call:
lm(formula = taux ~ traitement, data = exo1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42857 -1.07143 -0.07143  1.03571  3.42857
```

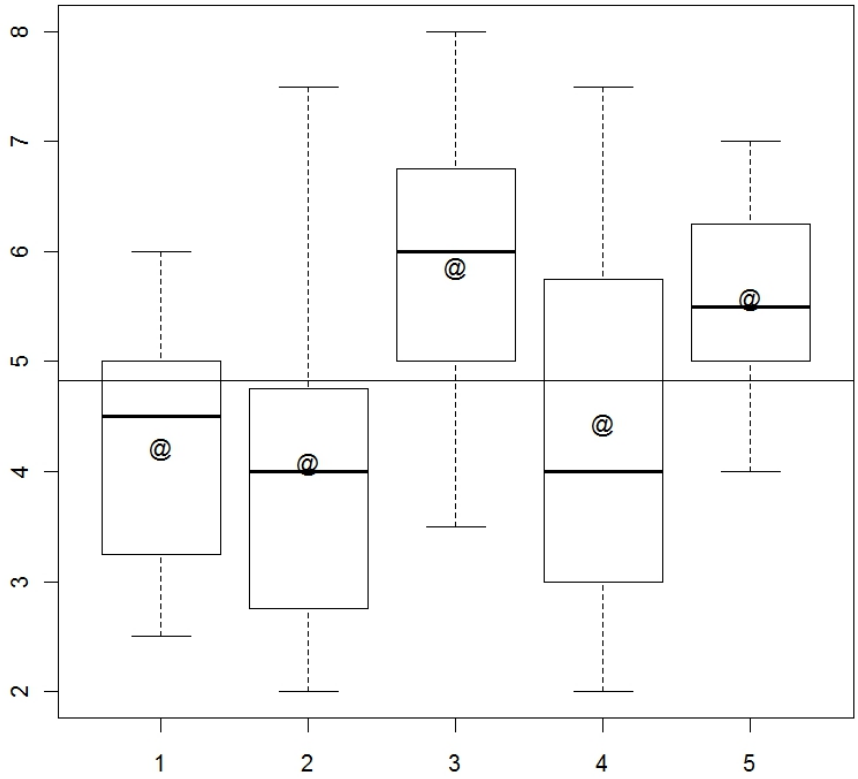

FIGURE 1. Les boîtes à moustaches pour les 5 traitements

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8286     0.2698  17.897   <2e-16 ***
traitement1  -0.6143     0.5396  -1.138   0.2639
traitement2  -0.7571     0.5396  -1.403   0.1708
traitement3   1.0286     0.5396   1.906   0.0662 .
traitement4  -0.4000     0.5396  -0.741   0.4643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.596 on 30 degrees of freedom
Multiple R-squared: 0.1995,     Adjusted R-squared: 0.09272
F-statistic: 1.869 on 4 and 30 DF,  p-value: 0.1419

> TukeyHSD(modele1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = taux ~ traitement, data = exo1)
```

```
$traitement
          diff        lwr      upr     p adj
2-1 -0.1428571 -2.6175554 2.331841 0.9998128
3-1  1.6428571 -0.8318411 4.117555 0.3262525
4-1  0.2142857 -2.2604125 2.688984 0.9990697
5-1  1.3571429 -1.1175554 3.831841 0.5145534
3-2  1.7857143 -0.6889839 4.260413 0.2493343
4-2  0.3571429 -2.1175554 2.831841 0.9932296
5-2  1.5000000 -0.9746982 3.974698 0.4156363
4-3 -1.4285714 -3.9032696 1.046127 0.4641681
5-3 -0.2857143 -2.7604125 2.188984 0.9971344
5-4  1.1428571 -1.3318411 3.617555 0.6693955
> TukeyHSD(modele2)
Erreur dans UseMethod("TukeyHSD") :
  pas de méthode pour 'TukeyHSD' applicable pour un objet de classe "lm"

> residus1<-residuals(modele1)
> residus1
          1           2           3           4           5           6
 0.28571429 -1.71428571  1.78571429  0.28571429 -1.21428571  1.28571429
          7           8           9          10          11          12
-0.71428571  3.42857143 -1.07142857 -1.57142857 -0.07142857 -2.07142857
         13          14          15          16          17          18
-0.07142857  1.42857143  2.14285714  0.64285714  0.14285714 -2.35714286
         19          20          21          22          23          24
-0.85714286  1.14285714 -0.85714286 -2.42857143  3.07142857 -0.42857143
         25          26          27          28          29          30
-1.92857143  0.57142857 -0.92857143  2.07142857  0.92857143 -0.07142857
         31          32          33          34          35
 0.42857143 -1.07142857 -1.57142857  1.42857143 -0.07142857
> shapiro.test(residus1)

        Shapiro-Wilk normality test

data:  residus
W = 0.9744, p-value = 0.5734
> residus2<-residuals(modele2)
> residus2
          1           2           3           4           5           6
 0.28571429 -1.71428571  1.78571429  0.28571429 -1.21428571  1.28571429
          7           8           9          10          11          12
-0.71428571  3.42857143 -1.07142857 -1.57142857 -0.07142857 -2.07142857
         13          14          15          16          17          18
-0.07142857  1.42857143  2.14285714  0.64285714  0.14285714 -2.35714286
         19          20          21          22          23          24
-0.85714286  1.14285714 -0.85714286 -2.42857143  3.07142857 -0.42857143
```

```
         25          26          27          28          29          30
-1.92857143  0.57142857 -0.92857143  2.07142857  0.92857143 -0.07142857
         31          32          33          34          35
 0.42857143 -1.07142857 -1.57142857  1.42857143 -0.07142857
> bartlett.test(residus1~traitement)

        Bartlett test of homogeneity of variances

data:  residus by traitement
Bartlett's K-squared = 3.1361, df = 4, p-value = 0.5353
> bartlett.test(residus2~traitement)

        Bartlett test of homogeneity of variances

data:  residus2 by traitement
Bartlett's K-squared = 3.1361, df = 4, p-value = 0.5353
```

Exercice 2. Quelles sont les variables qui influencent les ventes ?

Nous cherchons à étudier l'influence du marché total de la branche (MT), des remises aux grossistes (RG), des prix (PRIX), du budget de recherche (BR), des investissements (INV), de la publicité (PUB), des frais de ventes (FV) et du total du budget publicité de la branche (TPUB) sur les ventes semestrielles d'un certain produit. Toutes ces variables sont exprimées en K€. Les données concernant ces variables, pour 38 semestres, sont reportées dans le tableau présenté ci-dessous.

Semestre	MT	RG	PRIX	BR	INV	PUB	FV	TPUB	VENTES
1	398	138	56	12	50	77	229	98	5540
2	369	118	59	9	17	89	177	225	5439
3	268	129	57	29	89	51	166	263	4290
4	484	111	58	13	107	40	258	321	5502
5	394	146	59	13	143	52	209	407	4872
6	332	140	60	11	61	21	180	247	4708
7	336	136	60	25	-30	40	213	328	4627
8	383	104	60	21	-45	32	201	298	4110
9	285	105	63	8	-28	12	176	218	4123
10	277	135	62	11	76	68	175	410	4842
11	456	128	65	22	144	52	253	93	5741
12	355	131	65	24	113	77	208	307	5094
13	364	120	64	14	128	96	195	107	5383
14	320	147	66	15	10	48	154	305	4888
15	311	143	67	22	-25	27	181	60	4033
16	362	145	67	23	117	73	220	239	4942
17	408	131	66	13	120	62	235	141	5313
18	433	124	68	8	122	25	258	291	5014
19	359	106	69	27	71	74	196	414	5397
20	476	138	71	18	4	63	279	206	5149
21	415	148	69	8	47	29	207	80	5151
22	420	136	70	10	8	91	213	429	4989
23	536	111	73	27	128	74	296	273	5927
24	432	152	73	16	-50	16	245	309	4704
25	436	123	73	32	100	43	276	280	5366
26	415	119	75	20	-40	41	211	315	4630
27	462	112	73	15	68	93	283	212	5712
28	429	125	74	11	88	83	218	118	5095
29	517	142	74	27	27	75	307	345	6124
30	328	123	77	20	59	88	211	141	4787
31	418	135	79	35	142	74	270	83	5036
32	515	120	77	23	126	21	328	398	5288
33	412	149	78	36	30	26	258	124	4647

Semestre	MT	RG	PRIX	BR	INV	PUB	FV	TPUB	VENTES
34	455	126	78	22	18	95	233	118	5316
35	554	138	81	20	42	93	324	161	6180
36	441	120	80	16	-22	50	267	405	4801
37	417	120	81	35	148	83	257	111	5512
38	461	132	82	27	-18	91	267	170	5272

1. Écrire le modèle de régression permettant d'expliquer les ventes à l'aide de toutes les variables explicatives proposées. Préciser la nature de chacune des variables présentes dans le modèle ainsi que les conditions du modèle. Ces conditions sont-elles vérifiées ?
2. D'après les sorties statistiques réalisées avec le logiciel `R` qui se trouvent après, donner les estimations de tous les paramètres du modèle. Tester-les. Donner un intervalle de confiance à 95 % pour les paramètres du modèle.
3. En utilisant les résultats ci-après, donner le tableau de l'analyse de variance.
4. Ce modèle est-il intéressant ? Quelles sont les variables explicatives qui influent significativement sur le volume des ventes ? Est-il pertinent de simplifier le modèle introduit en 1. ? Si oui, expliquer celui que vous allez choisir pour optimiser la modélisation.
5. Dans cette question, on suppose que MT=500, RG=100, PRIX=83, BR=30, INV=50, PUB=90, FV= 300 et TPUB=200. Déterminer la valeur des ventes prédite par le modèle ? À l'aide des résultats ci-dessous, donner un intervalle de prédiction à 95 % pour cette valeur.

```
> exo2<-read.table(file.choose())
> str(exo2)
'data.frame':   38 obs. of  10 variables:
 $ Semestre : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MT       : int  398 369 268 484 394 332 336 383 285 277 ...
 $ RG       : int  138 118 129 111 146 140 136 104 105 135 ...
 $ PRIX     : int  56 59 57 58 59 60 60 60 63 62 ...
 $ BR       : int  12 9 29 13 13 11 25 21 8 11 ...
 $ INV      : int  50 17 89 107 143 61 -30 -45 -28 76 ...
 $ PUB      : int  77 89 51 40 52 21 40 32 12 68 ...
 $ FV       : int  229 177 166 258 209 180 213 201 176 175 ...
 $ TPUB     : int  98 225 263 321 407 247 328 298 218 410 ...
 $ VENTES   : int  5540 5439 4290 5502 4872 4708 4627 4110 4123 4842 ..
> modele23456789<-lm(VENTES~MT+RG+PRIX+BR+INV+PUB+FV+TPUB,data=exo2)
> residus23456789<-residuals(modele23456789)
> residus23456789
           1            2            3            4            5
 120.44004419  248.86870686 -228.65515795  -39.12445209 -415.52146821
           6            7            8            9           10
 176.27750551   75.79078136 -492.53754029  147.16415241  166.67831599
          11           12           13           14           15
```

```
 242.61366678  -46.46146912    1.81213981  384.46402277 -186.68944330
           16            17            18            19            20
-219.36746246    0.01994635  -88.65680627  468.52173238 -253.97984538
           21            22            23            24            25
 265.22668474 -349.75014334    6.56121013   56.83971851  236.53687528
           26            27            28            29            30
 -36.22494049   54.04485386 -297.77406256  411.54282381  -64.45623836
           31            32            33            34            35
-288.60923901  -77.98203272  -44.03029434 -110.58512418  176.75475941
           36            37            38
-107.44260419  174.91837801  -67.22799387
> shapiro.test(residus23456789)

        Shapiro-Wilk normality test

data:  residus
W = 0.988, p-value = 0.9504
> summary(modele23456789)

Call:
lm(formula = VENTES ~ MT + RG + PRIX + BR + INV + PUB + FV + TPUB, data = exo2)

Residuals:
    Min      1Q  Median      3Q     Max
-492.54 -109.80  -18.10  172.86  468.52

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3124.92373  643.10176   4.859 3.75e-05 ***
MT             4.50675    1.59254   2.830  0.00837 **
RG             1.78469    3.30031   0.541  0.59280
PRIX         -14.03885    8.32744  -1.686  0.10256
BR            -2.35462    6.58730  -0.357  0.72334
INV            1.83264    0.77987   2.350  0.02579 *
PUB            8.76775    1.83146   4.787 4.58e-05 ***
FV             1.33194    2.77815   0.479  0.63523
TPUB          -0.02725    0.40168  -0.068  0.94637
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 257 on 29 degrees of freedom
Multiple R-squared: 0.8048,     Adjusted R-squared: 0.751
F-statistic: 14.95 on 8 and 29 DF,  p-value: 2.097e-08
> anova(modele23456789)
Analysis of Variance Table

Response: VENTES
```

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
MT         1 5060415 5060415 76.6242 1.235e-09 ***
RG         1    9582    9582  0.1451 0.7060505
PRIX       1  153786  153786  2.3286 0.1378491
BR         1   12601   12601  0.1908 0.6654836
INV        1 1067509 1067509 16.1641 0.0003784 ***
PUB        1 1578406 1578406 23.9000 3.456e-05 ***
FV         1   14899   14899  0.2256 0.6383672
TPUB       1     304     304  0.0046 0.9463728
Residuals 29 1915218   66042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> modele2345678<-lm(VENTES~MT+RG+PRIX+BR+INV+PUB+FV,data=exo2)
Call:
lm(formula = VENTES ~ MT + RG + PRIX + BR + INV + PUB + FV , data = exo2)

Residuals:
    Min      1Q  Median      3Q     Max
-491.65 -110.68  -17.71  172.84  463.58

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3107.7194   581.1120   5.348 8.72e-06 ***
MT               4.5106     1.5649   2.882  0.00723 **
RG               1.8221     3.1996   0.569  0.57328
PRIX           -13.9359     8.0510  -1.731  0.09374 .
BR           -  2.3323     6.4690  -0.361  0.72098
INV              1.8406     0.7581   2.428  0.02140 *
PUB              8.7822     1.7886   4.910 3.00e-05 ***
FV               1.3125     2.7171   0.483  0.63257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 252.7 on 30 degrees of freedom
Multiple R-squared:  0.8048,    Adjusted R-squared:  0.7592
F-statistic: 17.67 on 7 and 30 DF,  p-value: 4.74e-09
> modele234678<-lm(VENTES~MT+RG+PRIX+INV+PUB+FV,data=exo2)
> summary(modele234678)

Call:
lm(formula = VENTES ~ MT + RG + PRIX + INV + PUB + FV, data = exo2)

Residuals:
    Min      1Q  Median      3Q     Max
-514.04 -111.01  -26.99  181.43  443.94

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3122.7770   571.4180   5.465 5.66e-06 ***
MT               4.7373     1.4127   3.353  0.00212 **
RG               1.8216     3.1544   0.577  0.56778
PRIX           -14.8606     7.5237  -1.975  0.05721 .
INV              1.8090     0.7424   2.437  0.02076 *
PUB              8.7550     1.7618   4.969 2.34e-05 ***
FV               0.9468     2.4850   0.381  0.70580
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 249.1 on 31 degrees of freedom
Multiple R-squared:  0.8039,    Adjusted R-squared:  0.766
F-statistic: 21.19 on 6 and 31 DF,  p-value: 1.034e-09
> modele23467<-lm(VENTES~MT+RG+PRIX+INV+PUB,data=exo2)
> summary(modele23467)

Call:
lm(formula = VENTES ~ MT + RG + PRIX + INV + PUB, data = exo2)

Residuals:
    Min      1Q  Median      3Q     Max
-516.22 -126.35  -28.53  181.57  432.80

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3069.4206   546.5414   5.616 3.31e-06 ***
MT               5.2017     0.7044   7.384 2.12e-08 ***
RG               1.8161     3.1119   0.584  0.56359
PRIX           -13.5425     6.5912  -2.055  0.04816 *
INV              1.9081     0.6860   2.782  0.00899 **
PUB              8.5649     1.6669   5.138 1.33e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 245.8 on 32 degrees of freedom
Multiple R-squared:  0.803,     Adjusted R-squared:  0.7723
F-statistic: 26.09 on 5 and 32 DF,  p-value: 2.034e-10
> modele2467<-lm(VENTES~MT+PRIX+INV+PUB,data=exo2)
> summary(modele2467)

Call:
lm(formula = VENTES ~ MT + PRIX + INV + PUB, data = exo2)

Residuals:
    Min      1Q  Median      3Q     Max
-564.30 -134.44  -39.32  186.25  442.29
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3302.3760   369.5658   8.936 2.50e-10 ***
MT               5.1695     0.6952   7.436 1.52e-08 ***
PRIX           -13.2269     6.5031  -2.034  0.05006 .
INV              1.8905     0.6784   2.787  0.00876 **
PUB              8.4596     1.6405   5.157 1.16e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 243.3 on 33 degrees of freedom
Multiple R-squared:  0.8009,    Adjusted R-squared:  0.7768
F-statistic: 33.19 on 4 and 33 DF,  p-value: 3.857e-11


> confint(modele12345678)
                    2.5 %       97.5 %
(Intercept) 1809.6329471 4440.2145165
MT             1.2496330    7.7638712
RG            -4.9652067    8.5345962
PRIX         -31.0703695    2.9926715
BR           -15.8271668   11.1179177
INV            0.2376256    3.4276593
PUB            5.0219839   12.5135067
FV            -4.3500188    7.0138995
TPUB          -0.8487831    0.7942771
> predict(modele12345678,data.frame(MT=500, RG=100, PRIX=83, BR=30, INV=50,
PUB=90, FV= 300, TPUB=200),interval="confidence")
       fit      lwr      upr
1 5595.767 5317.393 5874.141
> predict(modele12345678,data.frame(MT=500, RG=100, PRIX=83, BR=30, INV=50,
PUB=90, FV= 300, TPUB=200),interval="prediction")
       fit      lwr     upr
1 5595.767 5001.003 6190.53
```

Exercice 3. Le cancer de la gorge.

Nous souhaitons étudier la liaison entre les caractères : **être fumeur** (plus de 20 cigarettes par jour, pendant 10 ans) et **avoir un cancer de la gorge**, sur une population de 1000 personnes, dont 500 sont atteintes d'un cancer de la gorge. Voici les résultats observés :

Observé	**cancer**	**pas de cancer**
fumeur	342	258
non fumeur	158	242

Le fait de fumer plus de 20 cigarettes par jour, pendant 10 ans déclenche-t-il un cancer de la gorge ?
Pour répondre à la question, vous effectuerez un test dont vous donnerez le nom, puis vous énoncerez les deux hypothèses associées à ce test ainsi que la valeur de la statistique de ce test. Enfin, il manque deux valeurs notées `A` et `B` dans la sortie de `R`. Retrouvez-les.

```
> exo2<-matrix(c(342,258,158,242),
byrow=T,nrow=2,
dimnames=list(c("fumeur","non fumeur"),c("cancer","pas de cancer")))
> exo2
           cancer pas de cancer
fumeur        342           258
non fumeur    158           242

> chisq.test(exo2,correct=FALSE)

        Pearson's Chi-squared test

data:  exo2
X-squared = 29.4, df = 1, p-value = 5.888e-08

> chisq.test(exo2,correct=FALSE)$expected
           cancer pas de cancer
fumeur          A           300
non fumeur      B           200
```