INSA Méthodes Statistiques M8 Examen Final 2013 **- 3 h**

Sauf indication contraire, on considérera dans les exercices un risque de première espèce de 5%.

Exercice 1 — Zé doux, thé — 2 points

Au niveau national, le score moyen sur les examens d'entrée au collège est 452 avec un écart type de 95. Un échantillon aléatoire de 152 étudiants de première année entrant au collège de Touffreville montre un score moyen de 502.

1. Y a t-il une différence significative entre le collège de Touffreville et le niveau national ?

Solution de l'exercice 1

il faut faire un Z test (loi de Gauss)
z = -50/(95/sqrt(152)) z = -6.4889
pval = cdf('norm',z,0,1) pval = 4.3245e-11
on rejette H_0 : on peut dire qu'il y a une différence significative entre le collège de Touffreville et le niveau national

Exercice 2 — T'es où, Zed ? — 2 points

L'alcalinité, en milligrammes par litre, de l'eau dans le cours supérieur et inférieur des rivières dans une région donnée est connu pour être distribué normalement. Dix lectures d'alcalinité sont faits dans le cours supérieur d'une rivière dans la région et quinze dans le cours inférieur de la même rivière avec les résultats suivants.

cours supérieur	91 75 91 88 94 63 86 77 71 69
cours inférieur	86 95 135 121 68 64 113 108 79 62 143 108 121 85 97

1. Peut on affirmer, avec une risque de première espèce de 1%, que l'alcalinité de l'eau dans le cours inférieur de cette rivière est supérieure à celle dans la partie supérieure ?

Solution de l'exercice 2

Il faut faire un test de Student

```
 xi = [91 75 91 88 94 63 86 77 71 69];
xa = [ 86 95 135 121 68 64 113 108 79 62 143 108 121 85 97];
n1 = length(xi)
n1 = 10
n2 = length(xa)
n2 = 15
mi = mean(xi)
mi =   80.5000
ma = mean(xa)
ma =    99

sig = 1/(n1+n2-2)*(sum((xi-mi).^2) + sum((xa-ma).^2))
sig =  432.9783
t = mi-ma/sqrt(sig*(1/n1+1/n2))
t =   68.8459
pval = 1- cdf('t',t,n1+n2-2)
pval =     0
```

on peut donc affirmer, avec un risque de première espèce de 1%, que l'alcalinité de l'eau dans le cours inférieur de cette rivière est supérieure à celle dans la partie supérieure.

Exercice 3 — A Paris, hé ! — 3 points

Un aliment de commodité, connu comme « Quicknosh », a été introduit sur le marché en janvier 2010. Après une mauvaise année pour les ventes le fabricant a lancé une campagne de publicité intensive au cours de Janvier 2011. Le tableau ci-dessous enregistre les ventes, en milliers d'euros, pour une période d'un mois avant et d'un mois après la campagne de publicité, pour chacune des onze régions.

régions	nrd	sud	est	out	nes	ses	sse	nne	nno	sso	cen
ventes avant campagne	2,4	2,6	3,9	2,0	3,2	2,2	3,3	2,1	3,1	2,2	2,8
ventes après campagne	3,0	2,5	4,0	4,1	4,8	2,0	3,4	4,0	3,3	4,2	3,9

1. Peut on affirmer qu'une augmentation des ventes s'est produite ?

Solution de l'exercice 3

Il faut faire un test de Student sur données appariées

```
X = [2.4 3.0
2.6 2.5
3.9 4.0
2.0 4.1
3.2 4.8
2.2 2.0
3.3 3.4
2.1 4.0
3.1 3.3
2.2 4.2
2.8 3.9]

D = X(:,1) - X(:,2)
D =   -0.6    0.1   -0.1   -2.1  -1.6    0.2  -0.1   -1.9   -0.2   -2.0   -1.1

n = length(D)
n = 11
m = mean(D)
m =   -0.8545

sig = 1/(n-1)*sum((D-m).^2)
sig =    0.8227
t = m/sqrt(sig*(1/n))
t =   -3.1247

pval = cdf('t',t,n-1)
pval =   0.0054
```

on peut donc affirmer, avec un risque de première espèce de 5%, qu'une augmentation des ventes s'est produite.

Exercice 4 — Les kids de Cold play — 4 points

Des études de marché amènent à penser que la musique de fond pourrait affecter le comportement d'achat des clients. Une étude dans un supermarché a comparé trois situations affectées au hasard : pas de musique, musique d'accordéon français, et une musique de mandoline italienne. Dans chaque situation, le nombre de bouteilles de vin français, italien, et autre achetés a été enregistré. Voici un tableau qui résume les résultats :

	Musique		
	aucune	accordéon	mandoline
vin français	30	39	30
vin italien	11	5	19
autre	84	75	84

1. Peut on affirmer que le type de musique influence les ventes ?

Solution de l'exercice 4

Il faut faire un test du chi2

```
O = [30 39 30
11 5 19
84 75 84]

m = sum(O);
n = sum(O,2);

T = n*m/sum(n),
T =
   32.8249   31.2493   34.9257
   11.6048   11.0477   12.3475
   80.5703   76.7029   85.7268

D = sum(sum((O-T).^2./T))
D =   10.

ddl = (length(n)-1)*(length(m)-1)
ddl = 4
pval = 1 - chi2cdf(D,ddl)
pval =    0.0403
```

on peut donc affirmer, avec un risque de première espèce de 5%, que le type de musique influence les ventes.

Exercice 5 — L'art est graisse ion — 9 points

On a observé deux variables : la température (notée t) et le temps nécessaire au démarrage (notée d). Les observations ont été résumées dans le tableau suivant :

i	t	d
1	9.15	0.53
2	5.08	1.44
3	11.06	3.13
4	9.82	1.24
5	3.93	3.23
6	-0.93	3.94
7	8.18	0.16
8	-3.18	3.94
9	3.19	3.83
10	5.79	0.79
11	-2.05	3.96
12	1.20	3.90
13	7.37	0.03
14	-0.36	3.89
15	2.48	3.97

et on donne les résultats suivant :

$$\sum_{i=1}^{15} t_i = 60.77 \qquad \sum_{i=1}^{15} d_i = 38.07 \qquad \sum_{i=1}^{15} t_i^2 = 532.33 \qquad \sum_{i=1}^{15} d_i^2 = 132.70 \qquad \sum_{i=1}^{15} t_i d_i = 79.18$$

1. la régression linéaire simple :
   a) posez un modèle de régression linéaire de d en fonction de t
   pour chaque observation $i = 1, 15$: $d_i = at_i + b + \varepsilon_i$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
   b) estimez les paramètres du modèle
   on a

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{15} t_i = \frac{60.77}{15} = 4.05 \qquad \text{et} \qquad \bar{d} = \frac{1}{n}\sum_{i=1}^{15} d_i = \frac{38.07}{15} = 2.54$$

$$\text{d'où } \widehat{a} = \frac{n\sum_{i=1}^{n} d_i t_i - \sum_{i=1}^{n} d_i \sum_{i=1}^{n} t_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2} = \frac{15\times 79.18 - 60.77\times 38.07}{15\times 532.33 - 60.77^2} = -0.26 \text{ et}$$

$$\widehat{b} = \bar{d} - \widehat{a}\bar{t} = 2.54 + 0.26 \times 4.05 = 3.60$$

c) proposez une estimation de la variance de l'erreur
1.26

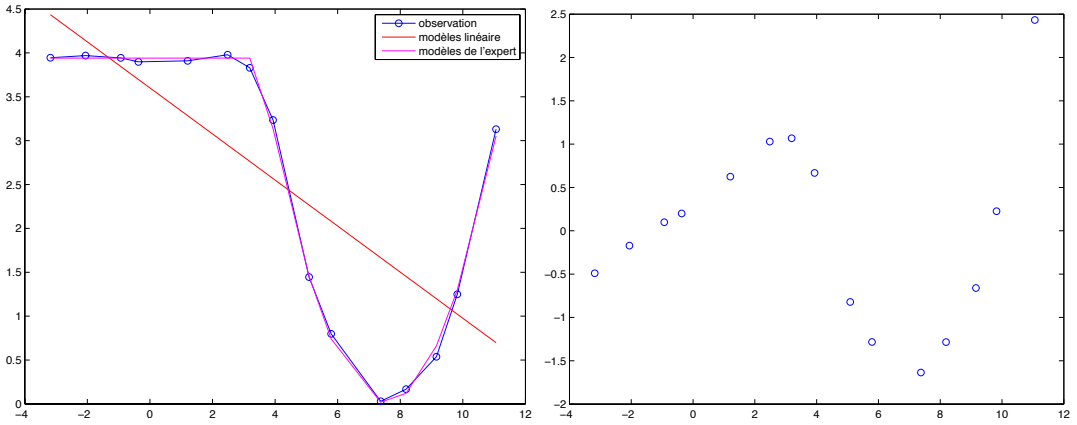
d) quelle serait selon vous le temps nécessaire au démarrage pour une température de -4
$\widehat{a}(-4) + \widehat{b} = 4.65$

e) donnez un intervalle de confiance sur cette prédiction.
on prend un student à 13 ddl avec $\alpha = 0.05$ on lit dans les tables $t_{\frac{\alpha}{2}} = 2,16$.

$$t_{\frac{\alpha}{2}}\ \widehat{\sigma}\ \sqrt{1 + \frac{1}{n} + \frac{(-4-\bar{t})^2}{\sum_{i=1}^{n}(t_i - \bar{t})^2}} = 2,16 \times \sqrt{4.65} \times \sqrt{1 + \frac{1}{15} + \frac{(-4-4.05)^2}{532.33 - 60.77^2}} = 5.30$$

on a donc une prévision dans l'intervalle $4.65 - 5.3 = -0.65; 4.65 + 5.3 = 9.95$

f) que pensez vous de la qualité du modèle ? Il est mauvais car les résidus sont structurés (cf figure ci dessous)



2. L'expert du domaine prétend que pour une température inférieure à 3,5, le temps nécessaire au démarrage est constant, et que ce temps est un polynôme du second degré en fonction de la température lorsque celle ci dépasse 3,5.

a) posez le modèle de régression associé aux dires de l'expert

$$f(t) = \begin{cases} a & \text{si } t \leq 3,5 \\ b + ct + dt^2 & \text{sinon} \end{cases}$$

Le modèle étant continu en 3.5 on a $a = b + c3.5 + d(3.5)^2$. ce qui donne le modèle réduit suivant

$$f(t) = \begin{cases} a & \text{si } t \leq 3,5 \\ c(t-3,5) + dt^2 - 12,25 & \text{sinon} \end{cases}$$

ou

$$f(t) = \begin{cases} b + 3.5c + 12.25d & \text{si } t \leq 3,5 \\ b + ct + dt^2 & \text{sinon} \end{cases}$$

b) écrire ce modèle sous forme matricielle : $d = Xa + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ et

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0.43 & 3.25 \\ 1 & 1.58 & 13.61 \\ 1 & 2.29 & 21.31 \\ 1 & 3.87 & 42.15 \\ 1 & 4.69 & 54.79 \\ 1 & 5.66 & 71.61 \\ 1 & 6.32 & 84.28 \\ 1 & 7.56 & 110.16 \end{pmatrix} \quad \text{ou } X = \begin{pmatrix} 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & 3.5 & 12.25 \\ 1 & -3.18 & 3.94 \\ 1 & 3.19 & 3.83 \\ 1 & 5.79 & 0.79 \\ 1 & -2.05 & 3.96 \\ 1 & 1.20 & 3.90 \\ 1 & 7.37 & 0.03 \\ 1 & -0.36 & 3.89 \\ 1 & 2.48 & 3.97 \end{pmatrix}$$

selon la variable que l'on décide d'éliminer.

c) proposez une fonction de type Matlab permettant d'estimer sous forme vectorielle les paramètres du modèle.
Voice un programe pour la première construction de la matrice X. Pour l'autre, seule la ligne correspondant à la constructyion de X change.

```
[t ind] = sort(t); % pour des raisons de commodité on trie les observations
d = d(ind)

ind = find(t>3.5); % recherche des valeurs inférieures à 3.5
n0 = length(find(t<=3.5));

X = [ones(n,1) [zeros(n0,1) t(ind)-3.5] [zeros(n0,1) t(ind)'.^2-3.5^2]];
%X = [ones(n,1) [3.5*ones(n0,1) t(ind)] [12.25*ones(n0,1) t(ind)'.^2]];
a = (X'*X)\(X'*y)
```

d) si ce modèle était exact, quelle serait selon vous le temps nécessaire au démarrage pour une température de -4 ?
le programme donne a = 3.94 -3.64 0.24. Le prédiction est donc la constante 3.94.