Sauf indication contraire, on considèrera dans les exercices un risque de première espèce de 5%. Pensez à vérifier la cohérence des données et à éliminer s'il y a lieu et en le justifiant les mesures hors épure.

Exercice 1 Zède ou thé 2 points

Une compagnie pharmaceutique veut savoir si le procédé de fabrication qu'elle utilise fournit effectivement des comprimés dosés à 20mg de principe actif d'un médicament. La quantité de principe actif est mesurée pour 16 comprimés issus d'un lot de fabrication, la mesure étant supposée suivre une loi normale. Le dosage moyen de principe actif est de 19,12 mg (la variance estimée étant de 3,11). La quantité moyenne de principe actif mesurée dans les comprimés de ce lot s'écarte-t-elle significativement du dosage prévu par le processus de fabrication ?

la moyenne suit une loi normale

$$\begin{aligned} H_0 : &\quad \text{la moyenne} = 20 \\ H_1 : &\quad \text{la moyenne} \neq 20 \end{aligned}$$

la variance étant estimée (donc inconue), la statistique du test

$$T = \frac{\overline{X}}{\sqrt{\frac{\widehat{\sigma}^2}{n}}}$$

suit une la loi de student à $n-1$ degrés de liberté

```
n = 16;
m = 19.12;
s2 = 3.11;
m0 = 20;

t = (m - m0)/(sqrt(s2/n))
pval = 2*cdf('t',t,n-1)
```

pour une valeur de $t = -1.9960$ on lit sur les table de la loi de student à $n-1 = 15$ degrés de libertè une $pval = 0.0644$, donc on garde H_0, la quantité moyenne de principe actif mesurée dans les comprimés de ce lot ne s'écarte pas significativement du dosage prévu par le processus de fabrication.

Exercice 2 Mauvais genre 2 points

On a classé les enfants d'un district écossais selon les modalités de deux caractères : le genre (garçon ou fille) et la couleur des cheveux : Blond, Roux, Châtain, Brun, Noir Jais. Le tableau de contingence suivant regroupe les résultats des observations :

	Blond	Roux	Châtain	Brun	Noir Jais
garçons	592	119	849	504	36
filles	544	97	677	451	14

A la vue de ce tableau, peut on affirmer que les caractères « genre » et « Couleur des cheveux » sont indépendants.

on fait un test de chi 2

$$\begin{aligned} H_0 : &\quad \text{indépendance} \\ H_1 : &\quad \text{dépendance} \end{aligned}$$

```
O = [592 119 849 504 36
544 97 677 451 14];

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

T = n*m/sum(n),
D = sum(sum((O-T).^2./T)) % 10.46

ddl = (length(n)-1)*(length(m)-1) % 4
pval = 1 - cdf('chi2',D,ddl) % 0.0332 < 0.05
```

pour une valeur de $t = -1.9960$ on lit sur les table de la loi du chi 2 à 4 degrés de libertè une $pval = 0.0332$, donc on rejette H_0, les caractères « genre » et « Couleur des cheveux » ne sont pas indépendants.

Exercice 3 Question boutique 2 points

Un échantillon aléatoire de 7 chocolats chauds du distributeur de Magellan avait une moyenne volume de 203 ml et un écart-type de 3 ml. Un autre échantillon aléatoire de 6 chocolats chauds du distributeur de Dumont D'urville a donné une moyenne de 208 ml et un écart-type de 5 ml. Peut on considérer, à la vue de cet échantillon, que les machines sont réglées de la même manière ?

on fait un test de student, les observations de Magellan suivant une loi normale d'espérance μ_1 et celles de Dumont D'urville une loi normale d'espérance μ_2

$$\begin{array}{ll} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{array}$$

```
n1 = 7;
m1 = 203;
s1 = 3^2;
n2 = 6;
m2 = 208;
s2 = 5^2;

SS = (1/n1 + 1/n2)*(n1*s1 + n2*s2)/(n1+n2-2);

T = abs(m1-m2)/sqrt(SS) % 2.04
pval = 2*(1-cdf('t',T,n1+n2-1)) % 0.06
```

pour une valeur de $t = 2.04$ on lit sur les table de la loi de student à 12 degrés de libertè une $pval = 0.06$, donc on garde H_0, on peut considérer, à la vue de cet échantillon, que les machines sont réglées de la même manière.

Exercice 4 La canice de jus de raisin 5 points

Le tableau suivant donne le revenu familial moyen en millier de dollars par habitant pour chacune des provinces du Canada et la consommation d'alcool en litres par personne de 15 ans et plus.

Province	Revenus	Alcool
Newfoundland	26.8	8.7
Prince Edward Island	27.1	8.4
Nova Scotia	29.5	8.8
New Brunswick	28.4	7.6
Quebec	30.8	8.9
Ontario	36.4	5.0
Manitoba	30.4	9.7
Saskatchewan	29.8	8.9
Alberta	35.1	11.1
British Columbia	32.5	10.9

1. Peut on affirmer que le coefficient directeur de la régression linéaire entre les variables Revenus (explicative) et Alcool (à expliquer) est égal à $\frac{1}{2}$?
2. Donnez une fourchette dans laquelle pourrait être la consommation moyenne d'alcool en litres par personne de 15 ans et plus de la province de Nunavut dont les revenus moyen seraient de 38 mille dollars.

```
data = [26.8 8.7
27.1 8.4
29.5 8.8
28.4 7.6
30.8 8.9
36.4 5.0
30.4 9.7
29.8 8.9
35.1 11.1
32.5 10.9];

n = length(data)
x = data(:,1);
X = [x ones(n,1)];
y = data(:,2);

a = X\y

e = y-X*a;
s2 = e'*e/(n-2);
R2 = 1 - e'*e/sum((y-mean(y)).^2);
h = diag(X*((X'*X)\(X')));
c = h./(1-h).^2/2.*e.^2/(e'*e);

dv = [e h c]
```

on élimine le point 6 qui est hors épure et qui a un résidu et une contribution trop importants

pour décider si le coefficient directeur de la régression linéaire entre les variables Revenus (explicative) et Alcool (à expliquer) est égal à $\frac{1}{2}$, nous allons faire une régression linéaire et faire un test de student sur notre estimation $\widehat{a}$ du coéfficient directeur

$$\begin{array}{ll} H_0: & \widehat{a} = \frac{1}{2} \\ H_1: & \widehat{a} \neq \frac{1}{2} \end{array}$$

```
data(6,:) = [];

x = data(:,1);
n = length(x);
X = [x ones(n,1)];
y = data(:,2);
a = X\y

e = y-X*a;
s2 = e'*e/(n-2);
R2 = 1 - e'*e/sum((y-mean(y)).^2);
h = diag(X*((X'*X)\(X')));
c = h./(1-h).^2/2.*e.^2/(e'*e);

dv = [e h c]

mx = mean(x);
s2x = (x-mx)'*(x-mx)
T = abs(a(1)-.5)/sqrt(s2/s2x)
pval = 2*(1-cdf('t',abs(T),n-2))
```

Pour une valeur de $t = 1.44$ on lit sur les table de la loi de student à $n-2=7$ degrés de libertè une $pval = 0.19$, on garde H_0, on peut affirmer que le coefficient directeur de la régression linéaire entre les variables Revenus (explicative) et Alcool (à expliquer) est égal à $\frac{1}{2}$.

Donnez une fourchette dans laquelle pourrait être la consommation moyenne d'alcool en litres par personne de 15 ans et plus de la province de Nunavut dont les revenus moyen seraient de 38 mille dollars.

```
xp = [38 1]
yp = xp*a
t = icdf('t',0.025,n-2);
interval = t*sqrt(s2*(1+1/n+(38-mean(x)).^2/s2x));
[yp-interval;yp+interval]
```

entre 10.48 et 13.98

Exercice 5 Il a parié ! 4 points

Une étude a pour but de comparer la perception des cadres supérieurs et des employés quant à l'autoritarisme de la direction des entreprises. On tire aléatoirement 10 entreprises et dans chacune d'elles on sélectionne au hasard, un cadre et un employé. Ils répondent chacun à un questionnaire qui permet de calculer un score global qui reflète leur perception du caractère autoritaire de la direction de leur entreprise. Les résultats sont les suivants :

entreprise	1	2	3	4	5	6	7	8	9	10
score de cadre	4.92	11.35	5.60	9.83	11.83	13.37	12.42	12.95	8.58	13.29
score de l'employé	4.59	10.82	5.87	9.97	11.71	12.53	10.09	12.51	17.48	12.85

Peut on affirmer (ou parier) que la perception de l'autoritarisme de la direction est plus importante chez les cadres que chez les employés ?

On fait un test de student apparié

$$\begin{array}{ll} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \geq \mu_2 \end{array}$$

```
O = [4.92 11.35 5.60 9.83 11.83 13.37 12.42 12.95 8.58 13.29
4.59 10.82 5.87 9.97 11.71 12.53 10.09 12.51 17.48 12.85]

d = (O(1,:)-O(2,:))';

plot(d,'o')
boxplot(d);
```

on élimine la mesure 9 aberrante. La mesure 7 est aussi sujette à caution...
on fait un test de student apparié

```
d(9) = [];
md = mean(d);
n = length(d);
s2d = (d-md)'*(d-md)/(n-1)

t = (md)/sqrt(s2d/n)
pval = (1-cdf('t',t,n-1))
```

pour une valeur de $t = 2.02$ on lit sur les table de la loi de student à $n - 1 = 8$ degrés de libertè une $pval = 0.039$, et donc on rejette H_0, la perception de l'autoritarisme de la direction est plus importante chez les cadres que chez les employés.

Exercice 6 le MCP 5 points

Les observation de deux variables x et y ont données les résultats suivants :

x	y
0	-0.0778
0.1	0.0935
0.2	0.2232
0.3	0.5577
0.4	0.6947
0.5	1.0648
0.6	1.1682
0.7	1.5769
0.8	1.7511
0.9	1.8164
1.0	1.9717
1.1	2.3052
1.2	2.2453
1.3	2.5774
1.4	2.5731
1.5	2.7100
1.6	2.7728
1.7	2.9676
1.8	2.8822
1.9	2.8836
2.0	3.0553
2.1	2.9823
2.2	3.0134

1. Proposez une méthode permettant d'estimer les paramètres du modèle continu suivant :

$$y_i = \varepsilon_i + \begin{cases} ax_i & \text{si } x_i \leq 1 \\ b + cx_i + dx_i^2 & \text{si } x_i > 1 \\ e & \text{si } x_i > 2 \end{cases}$$

2. En utilisant cette méthode, proposez une estimation des paramètres a, b, c, d et e.
3. Représentez graphiquement, avec les données y en fonction de x, la fonction $f(t)$

$$f(t) = \begin{cases} ax & \text{si } t \leq 1 \\ b + ct + dt^2 & \text{si } t > 1 \\ e & \text{si } t > 2 \end{cases}$$

pour $t \in [0, 3]$.

on a les équations de continuité au point $t = 1$ $a = b + c + d$ et au point $t = 2$ $b + 2c + 4d = e$. Le modèle linéaire s'écrit donc

$$\begin{cases} y_i = bx_i + cx_i + dx_i + \varepsilon_i & \text{si } x_i \leq 1 \\ y_i = b + cx_i + dx_i^2 + \varepsilon_i & \text{si } x_i > 1 \\ y_i = b + 2c + 4d + \varepsilon_i & \text{si } x_i > 2 \end{cases}$$

et donc de la forme

$$\mathbf{y} = X\alpha + \varepsilon$$

avec

$$X = \begin{cases} x_i \quad x_i \quad x_i & \text{si } x_i \leq 1 \\ 1 \quad x_i \quad x_i^2 & \text{si } 1 < x_i \leq 2 \\ 1 \quad 2 \quad 4 & \text{si } x_i > 2 \end{cases} \qquad \alpha = \begin{pmatrix} b \\ c \\ d \end{pmatrix}$$

La méhode est donc la suivante :

- on estime b, c et d par la méthode des moindres carrés en résolavant $\alpha = (X'X)^{-1}(X' * y)$
- puis on a $a = b + c + d$ et $e = b + 2c + 4d$

```
data = [0 -0.0778
0.1 0.0935
0.2 0.2232
0.3 0.5577
0.4 0.6947
0.5 1.0648
0.6 1.1682
0.7 1.5769
0.8 1.7511
0.9 1.8164
1.0 1.9717
1.1 2.3052
1.2 2.2453
1.3 2.5774
1.4 2.5731
1.5 2.7100
1.6 2.7728
1.7 2.9676
1.8 2.8822
1.9 2.8836
2.0 3.0553
2.1 2.9823
2.2 3.0134]

x = data(:,1);
n = length(x);
y = data(:,2);

x1 = x(1:11)
x2 = x(12:21);
un = ones(2,1);
deux = 2*un;
quatre = 4*un;
X = [x1 x1 x1; ones(size(x2)) x2 x2.^2 ; un deux quatre];

al = (X'*X)\(X'*y);

b = al(1)
c = al(2)
d = al(3)
a = b+c+d
e = b+2*c+4*d

t = 0:0.01:3;
ind1 = find(t<=1);
ind2 = find(t<=2 & t>1);
ind3 = find(t>2);
ym = [a*t(ind1) b+c*t(ind2)+d*t(ind2).^2 e*ones(size(ind3))]
```

on trouve $a = 2.05$, $b = -0.41$, $c = 3.21$, $d = -0.75$ et $e = 3.00$.

```
plot(t,ym);
hold on
plot(x,y,'or')
hold off
```