Bordeaux I, 2011-2012. M1 Enseignement des Mathématiques.

Probabilités et Statistiques

DS du 3 novembre 2011 (9h-12h)

Exercice 1

Compléter la fonction matlab ci-dessous pour qu'elle permette de simuler une variable aléatoire de densité $\frac{1}{x\ln(2)}1_{[1,2]}(x)$.

```
function X=toto()
 U=rand(1);
 ....
end;
```

Exercice 2

On suppose que pour tout entier $n \geq 1$, la probabilité qu'il y ait n enfants dans une famille est αp^n, où α et p sont deux paramètres réels positifs vérifiant $0 < p < 1$ et $0 < \alpha < \frac{1-p}{p}$. On suppose de plus que les sexes des différents enfants sont indépendants les uns des autres, et équiprobables. Les termes "Garçon" et "Fille" feront toujours référence aux enfants (et pas aux parents).

(1) Quelle est la probabilité qu'il n'y ait pas d'enfants dans une famille ?

(2) Soient m et k deux entiers fixés. On considère une famille de m enfants. Quelle est la probabilité dans ce modèle qu'elle comporte exactement k garçons ?

(3) Quelle est la probabilité qu'une famille comporte exactement k garçons pour $k = 0$? Pour $k = 1$? (Le nombre d'enfants de la famille n'est plus supposé connu égal à m).

(4) On considère une famille dont on sait qu'elle comporte au moins une fille.

   (a) Quelle est la probabilité qu'elle comporte exactement une fille (et un nombre quelconque de garçons) ?

   (b) Quelle est la probabilité qu'elle ne comporte pas de garçon ?

   (c) Quelle est la probabilité que cette famille ne comporte qu'un enfant ?

Problème

Question préliminaire.

Dans cette question, toutes les variables aléatoires sont supposées définies sur un même espace de probabilité et posséder une espérance et une variance.

Soit X et Y deux variables aléatoires. On définit la covariance de X et Y par

$$cov(X,Y) = E(\,(X - E(X))\,(Y - E(Y))\,)$$

On remarque l'égalité $cov(X,X) = var(X)$.

a) Montrer l'égalité

$$var(X+Y) = var(X) + var(Y) + 2\,cov(X,Y)$$

b) Montrer que si c_1 et c_2 sont deux réels, et si Y_1 et Y_2 sont deux variables aléatoires

$$cov(X, c_1 Y_1 + c_2 Y_2) = c_1\,cov(X,Y_1) + c_2\,cov(X,Y_2)$$

Vif du sujet.

Soit n un entier strictement positif, et (x_i, y_i) $(i = 1, ..., n)$ n couples de réels. On suppose que les x_i ne sont pas tous égaux. On pose

$$\overline{x} = \frac{\sum_{i=1}^n x_i}{n} \qquad \overline{y} = \frac{\sum_{i=1}^n y_i}{n}$$

On pourra remarquer l'égalité $\sum_{i=1}^n (x_i - \overline{x}) = 0$.

I- Droite des moindres carrés.

Pour a et b réels, on pose

$$h(a,b) = \sum_{i=1}^n (y_i - a x_i - b)^2$$

On remarque que $h(a,b) = 0$ si et seulement si les points (x_i, y_i) sont sur la droite d'équation $y = ax + b$.

De manière générale, la droite d'équation $y = ax + bx$ obtenue pour la valeur de (a,b) qui minimise $h(a,b)$ est appellée la *droite des moindres carrés.*

1) Soit a et b donnés. Montrer l'égalité

$$h(a,b) = n\,(\overline{y} - a\overline{x} - b)^2 \;+ \sum_{i=1}^n (\,(y_i - \overline{y}) - a(x_i - \overline{x})\,)^2$$

2) Pour quelle valeur de a la somme $\sum_{i=1}^n (\,(y_i - \overline{y}) - a(x_i - \overline{x})\,)^2$ est-elle minimale ?

3) Montrer que la droite des moindres carrés passe nécessairement par le point de coordonnées $(\overline{x}, \overline{y})$.

4) Quel est le couple (a, b) qui réalise le minimum de $h(a, b)$?

II- Modèle de régression linéaire.

On considère désormais que les valeurs des réels x_i $(i = 1, ..., n)$ ne sont pas aléatoires et qu'elles nous sont connues, et que les valeurs y_i sont les valeurs observées de variables aléatoires $Y_i = \alpha x_i + \beta + \varepsilon_i$, où

- α et β sont des réels fixés, mais qui nous sont inconnus,

- les ε_i $(i = 1, ..., n)$ sont des variables aléatoires indépendantes de même loi, d'espérance nulle et de variance σ^2 fixée, mais qui nous est inconnue. On note $\overline{\varepsilon} = (\sum_{i=1}^n \varepsilon_i)/n$.

On considère les statistiques suivantes

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad A = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad B = \overline{Y} - A\,\overline{x} \qquad H = \sum_{i=1}^n (Y_i - Ax_i - B)^2$$

1) a) Montrer l'égalité

$$A = \alpha + \frac{\sum_{i=1}^n (x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

b) Déterminer l'espérance et la variance de A.

c) L'estimateur A de α est-il sans biais ? Quelle est son erreur quadratique ?

2) a) Pour $i \in \{1, ..., n\}$, calculer $cov(\varepsilon_i, \overline{\varepsilon})$. En déduire que la covariance de A et $\overline{\varepsilon}$ est nulle.

b) Calculer la variance de $\overline{\varepsilon}$.

c) Montrer l'égalité

$$B = -(A - \alpha)\overline{x} + \beta + \overline{\varepsilon}$$

d) Déterminer l'espérance et la variance de B.

3) a) Montrer l'égalité

$$H = \sum_{i=1}^n (-(A - \alpha)(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})\,)^2$$

b) Soit i dans $\{1, ..., n\}$. Déterminer $E(\,(\varepsilon_i - \overline{\varepsilon})^2\,)$ et $E(\,\varepsilon_i(A - \alpha)\,)$.

c) Montrer que $H/(n-2)$ est un estimateur sans biais de σ^2.

III- Prévision à l'aide du modèle de régression linéaire.

On désigne par a et b les coefficients de la droite des moindres carrés définie dans le I à partir des données (x_i, y_i).

On se donne un réel x°. On cherche à estimer $\alpha x^\circ + \beta$ à l'aide du modèle de régression linéaire. Pour cela, on utilise l'estimateur $Y^\circ = Ax^\circ + B$, dont $y^\circ = ax^\circ + b$ est la valeur observée.

On suppose dans cette partie que la loi des variables aléatoires indépendantes ε_i est normale $\mathcal{N}(0, \sigma^2)$.

1) a) Expliquer pourquoi la loi de Y° est normale.

b) Quelle est l'espérance de Y° ?

c) Utiliser II-2-a et II-2-c pour montrer l'égalité $cov(A, B) = -\overline{x}\, var(A)$

d) Montrer que la variance de Y° vaut

$$\sigma^2(\frac{1}{n} + \frac{(x^\circ - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2})$$

e) Donner la loi de

$$\frac{Y^\circ - \alpha x^\circ - \beta}{\sigma\sqrt{\frac{1}{n} + \frac{(x^\circ - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

f) On admet que la loi de

$$\frac{Y^\circ - \alpha x^\circ - \beta}{\sqrt{\frac{H}{n-2}}\sqrt{\frac{1}{n} + \frac{(x^\circ - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}$$

est la loi de Student à $n-2$ degrés de liberté.

On suppose $n = 19$. Déterminer un intervalle de confiance au niveau de confiance de 0.95 pour l'estimation de $\alpha x^\circ + \beta$.

2) *Application*

Les températures moyennes annuelles à Mérignac pour les années 1991 à 2009 ($x_i = -9, ..., -1, 0, 1, .., 9$) ont été :

année $(2000 + x_i)$	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
température en d°C (y_i)	13.3	13.4	13.3	14.5	14.4	13.6	14.9	13.9	14.3	14.4

année $(2000 + x_i)$	2001	2002	2003	2004	2005	2006	2007	2008	2009
température en d°C (y_i)	13.9	14.4	14.9	13.8	13.8	14.5	13.8	13.7	14.2

Calculer les coefficients a et b de l'équation de la droite des moindres carrés.

En utilisant le modèle de régression linéaire, déterminer l'estimation de la température en 2020 ($x^\circ = 20$), et donner l'intervalle de confiance au niveau de confiance de 0.95 pour cette prédiction.

On utilisera les résultats de calculs suivants :

$\overline{y} \simeq 14.05 \quad \sum_{i=1}^{n}(x_i - \overline{x})^2 = 570 \quad \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 12.4$

$\sum_{i=1}^{n}(y_i - a x_i - b)^2 \simeq 4.14$

A.3 Loi de Student $\mathcal{T}(d)$

Graphe de la densité $\phi(t) = \frac{\Gamma((d+1)/2)}{\Gamma(d/2)\sqrt{\pi d}}(1+\frac{t^2}{d})^{-\frac{d+1}{2}}$

$X = \dfrac{U}{\sqrt{V/d}}$ où U et V suivent les lois $N(0,1)$ et $\chi^2(d)$ et sont indépendantes.

$\mathbb{E}(X) = 0$, $\text{Var}(X) = d/(d-2)$.

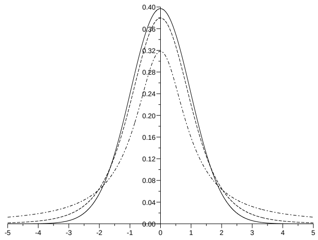

Table de dépassement de l'écart absolu : $\mathbb{P}(|X| > t_\alpha) = \alpha$

La première colonne donne le nombre de degrés de liberté *ddl*. La première ligne donne la probabilité α d'être dépassée. Par exemple, si $ddl = 10$ et $\alpha = 0.05$ alors $t_\alpha = 2.228$.

	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
45	0.680	1.301	1.679	2.014	2.412	2.690	2.952	3.281	3.520
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496