Université de Marseille
Licence de Mathématiques, 3eme année, probabilités-Statistique
Partiel du 16 mai 2017

Le polycopié du cours, les notes de cours et les notes de TD sont autorisés.
Le partiel contient 3 exercices. Le barème est sur 25 points.

ATTENTION ERRATUM, il y a eu des erreurs dans les données numériques de l'exercice 3 :
les données sont de taille 9 (et non 10) et les données concernant la vraisemblance sont erronnées, voir les commentaires dans le corrigé ci dessous.

Exercice 1 (Loi d'un produit, barème 5 points)
Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé, et X, Y deux v.a.r.. On suppose que la loi du vecteur aléatoire de composantes X, Y est une loi de densité f par rapport à la mesure de Lebesgue et que f est donnée par

$$f(x,y) = 10\,xy^2 \text{ si } 0 < x < y < 1,$$
$$f(x,y) = 0 \text{ sinon.}$$

1. Donner les densités (par rapport à la mesure de Lebesgue) des v.a.r. X et Y.

Corrigé – *On note g la densité de X et h la densité de Y. On a (pour presque tout $x \in \mathbb{R}$ et presque tout $y \in \mathbb{R}$) $g(x) = \int_{\mathbb{R}} f(x,y)dy$ et $h(y) = \int_{\mathbb{R}} f(x,y)dx$. Ceci donne (p.p. sur $\mathbb{R}$)*

$$g(x) = 10\,x \int_x^1 y^2 dy = \frac{10}{3} x(1-x^3) \text{ si } x \in]0,1[,$$
$$g(x) = 0 \text{ si } x \notin]0,1[.$$

$$h(y) = 10\,y^2 \int_0^y x dx = 5y^4 \text{ si } y \in]0,1[,$$
$$h(y) = 0 \text{ si } y \notin]0,1[.$$

2. Les v.a.r. X et Y sont-elles indépendantes ?

Corrigé – *Les v.a.r. X et Y ne sont pas indépendantes car $f(x,y) - g(x)h(y) \neq 0$ sur un ensemble de mesure non nulle de $\mathbb{R}^2$. Par exemple, $f(x,y) - g(x)h(y) \neq 0$ p.p. sur $\{(x,y) \in \mathbb{R}^2,\ 0 < y < x < 1\}$.*

3. Montrer que la v.a.r. XY a une densité par rapport à la mesure de Lebesque (sur les boréliens de $\mathbb{R}$) et donner une expression de cette densité.

Corrigé – *On pose $Z = XY$. Soit φ une fonction borélienne bornée de $\mathbb{R}$ da,s $\mathbb{R}$. On a, en utilisant le théorème de Fubini et un changement de variable,*

$$E(\varphi(Z)) = \int_{\mathbb{R}^2} \varphi(xy) f(x,y) d(x,y) = 10 \int_0^1 \left(\int_0^y xy\varphi(xy) y dx \right) dy = 10 \int_0^1 \int_0^{y^2} z\varphi(z) dz =$$
$$10 \int_0^1 \left(\int_{\sqrt{z}}^1 dy \right) z\varphi(z) dz = 10 \int_0^1 z(1-\sqrt{z})\varphi(z) dz.$$

Ceci prouve que la v.a.r. Z a une densité et cette densité est donnée par la fonction q définie (p.p. sur $\mathbb{R}$) par

$$q(x)) = 10\,x(1-\sqrt{x}) \text{ si } x \in]0,1[,$$
$$q(x) = 0 \text{ si } x \notin]0,1[.$$

Exercice 2 (Variations sur une suite de v.a.r.i.i.d., barème 10 points) Soientt $(\Omega, \mathcal{A}, P)$ un espace probabilisé et $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes suivant la loi $\mathcal{N}(0,1)$. Pour tout entier $n \geq 1$ on pose

$$Y_n = X_n + X_{n+1}.$$

1. Montrer que les variables aléatoires $(Y_n)_{n\geq 1}$ sont de même loi. Préciser cette loi.

Corrigé – *Soit $n \in \mathbb{N}^\star$. Comme X_n et X_{n+1} sont indépendantes, la loi de Y_n ne dépend que des lois de X_n et X_{n+1}. Comme tous les X_n ont même loi, la loi de Y_n ne dépend pas de n.*

Pour calculer la loi de Y_n, le plus rapide est probablement d'utiliser la fonction caractéristique d'une v.a.r. (la fonction caractéristique de Z est notée ci dessous ϕ_Z). Comme X_n et X_{n+1} sont indépendantes et de même loi, on a $\phi_{Y_n} = \phi_{X_n}\phi_{X_{n+1}} = \phi_{X_n}^2$. Comme $X_n \sim \mathcal{N}(0,1)$, on a, pour tout $t \in \mathbb{R}$, $\phi_{X_n}(t) = e^{-t^2/2}$. On a donc $\phi_{Y_n}(t) = e^{-t^2}$, ce qui prouve que $Y_n \sim \mathcal{N}(0,2)$.

Une autre méthode consiste à calculer $E(\varphi(Y_n))$ pour tout fonction φ une fonction borélienne bornée de $\mathbb{R}$ dans $\mathbb{R}$. On a, en utilisant les lois de X_n et X_{n+1}, le fait que X_n et X_{n+1} sont indépendantes, le théorème de Fubini et un changement de variable,

$$(2\pi)E(\varphi(Y_n)) = \int_{\mathbb{R}^2} \varphi(x+y)e^{-\frac{x^2}{2}}e^{-\frac{y^2}{2}}d(x,y) = \int_{\mathbb{R}} e^{-\frac{y^2}{2}}\left(\int_{\mathbb{R}} \varphi(x+y)e^{-\frac{x^2}{2}}dx\right)dy =$$
$$\int_{\mathbb{R}} e^{-\frac{y^2}{2}}\left(\int_{\mathbb{R}} \varphi(z)e^{-\frac{(y-z)^2}{2}}dz\right)dy = \int_{\mathbb{R}} e^{-y^2}\left(\int_{\mathbb{R}} \varphi(z)e^{-\frac{z^2}{2}}e^{yz}dz\right)dy =$$
$$\int_{\mathbb{R}} e^{-\frac{z^2}{2}}\varphi(z)\left(\int_{\mathbb{R}} e^{-y^2}e^{yz}dy\right)dz = \int_{\mathbb{R}} e^{-\frac{z^2}{4}}\varphi(z)\left(\int_{\mathbb{R}} e^{-(y-\frac{z}{2})^2}dy\right)dz = \sqrt{\pi}\int_{\mathbb{R}} e^{-\frac{z^2}{4}}\varphi(z)dz.$$

On a donc $E(\varphi(Y_n)) = \displaystyle\int_{\mathbb{R}} \frac{1}{\sqrt{4\pi}}e^{-\frac{z^2}{4}}\varphi(z)dz$.

Ceci prouve aussi que $Y_n \sim \mathcal{N}(0,2)$.

2. Pour $n \geq 1$ calculer $\mathbb{E}(Y_n)$, Var(Y_n) et Cov(Y_n, Y_{n+1}).

Corrigé – *Soit $n \in \mathbb{N}^\star$. Comme $Y_n \sim \mathcal{N}(0,2)$, on a $E(Y_n) = 0$, Var$(Y_n) = 2$. Puis en utilisant l'indépendance des X_n,*

$$Cov(Y_n, Y_{n+1}) = E(X_n + X_{n+1}, X_{n+1} + X_{n+2}) = E(X_{n+1}^2) = Var(X_{n+1}) = 1.$$

3. La suite $(Y_n)_{n\geq 1}$ est-elle une suite de variables aléatoires indépendantes ?

Corrigé – *La suite $(Y_n)_{n\geq 1}$ n'est pas une suite de variables aléatoires indépendantes. Pour le montrer, il suffit de remarquer que (pour tout n) $E(Y_nY_{n+1}) = Cov(Y_n, Y_{n+1}) = 1 \neq 0 = E(Y_n)E(Y_{n+1})$.*

4. La suite $(Y_{2n})_{n\geq 1}$ est-elle une suite de variables aléatoires indépendantes ?

Corrigé – *La suite $(Y_{2n})_{n\geq 1}$ est une suite de variables aléatoires indépendantes car elle est construite avec une fonction borélienne (la fonction, de $\mathbb{R}^2$ dans $\mathbb{R}$, $(x,y) \mapsto x+y$) et une suite de parties disjointes deux à deux de la suite de v.a.r. indéndantes $(X_n)_{n\geq 1}$ (les éléments de cette suite de parties sont les parties $\{X_{2n}, X_{2n+1}\}$ pour $n \in \mathbb{N}^\star$).*

5. Montrer que la suite $\left(\frac{1}{n}\sum_{k=1}^n Y_{2k}\right)_{n\geq 1}$ converge presque sûrement vers 0. On précisera et vérifiera avec soin les hypothèses du théorème utilisé.

Corrigé – *La suite $\left(\frac{1}{n}\sum_{k=1}^n Y_{2k}\right)_{n\geq 1}$ est une suite de v.a.r.i.i.d. intégrables. On peut donc comme appliquer la loi forte de grands nombres. Comme $E(Y_1) = 0$, on obtient bien le résultat demandé.*

6. On pose $M_n = \frac{1}{n}\sum_{k=1}^n Y_k$. Montrer que la suite $(M_n)_{n\geq 1}$ converge presque sûrement vers 0 (Indication : regrouper astucieusement les termes).

Corrigé – *On a $M_n = \frac{X_1}{n} + \frac{2}{n}\sum_{k=1}^n X_k + \frac{X_{n+1}}{n} = \frac{1}{n}\sum_{k=1}^n X_k + \frac{1}{n}\sum_{k=2}^{n+1} X_k$. La loi forte des grands nombres appliquée aux suites $(X_n)_{n\geq 1}$ et $(X_n)_{n\geq 2}$ donne que $\frac{1}{n}\sum_{k=1}^n X_k \to 0$ p.s. et $\frac{1}{n}\sum_{k=2}^{n+1} X_k \to 0$ p.s.. On en déduit que $M_n \to 0$ p.s..*

7. Pour $\omega \in \Omega$, on pose

$$Z(\omega) = \sum_{n=1}^{\infty} \left(\frac{Y_n(\omega)}{n} \right)^2 .$$

La variable aléatoire Z prend donc ses valeurs dans $\mathbb{R}_+ \cup \{+\infty\}$. Montrer que $\mathbb{E}(Z) < +\infty$.
En déduire que la suite $\left(\frac{Y_n}{n}\right)_{n\geq 1}$ tend presque sûrement vers 0.

Corrigé – Le théorème de convergence monotone donne que $E(Z) = \sum_{n=1}^{+\infty}(1/n^2)E(Y_n^2) = \sum_{n=1}^{+\infty}(2/n^2) < +\infty$. *On en déduit que* $|Z| < +\infty$ *p.s. et donc que* $Y_n/n \to 0$ *p.s..*

Exercice 3 (Statistique, barème 10 points) On dispose de données issues de 9 expériences concernant un caractère quantitatif aléatoire. Par la suite on notera ces données comme le vecteur x indiqué ci-dessous (où ${}^t()$ désigne la transposée) :

$$x =^t (4.52, 5.72, 4.47, 5.21, 4.71, 4.73, 4.88, 4.50, 5.80).$$

On utilise une approche de statistique inférentielle. Ainsi x est une réalisation d'un vecteur aléatoire noté X et dont on note X_i les composantes :

$$X =^t (X_1, X_2, \ldots, X_9).$$

On sait d'expérience qu'un bon modèle pour ce type de données consiste à supposer que les variables $X_1, \ldots, X_9$ sont, à valeurs réelles, indépendantes, et identiquement distribuées (iid) de loi commune donnée par :

$$\forall y \in \mathbb{R},\ q_\theta(y) = \frac{3}{4}\left(1 - (y-\theta)^2\right) \mathbb{I}_{[\theta-1,\theta+1]}(y).$$

Ici $\theta \in \mathbb{R}_+^*$ est un paramètre déterministe inconnu dont on cherche à deviner la valeur.

Rappel : ci dessus $\mathbb{I}$ *est la fonction indicatrice, dont on rappelle la définition : si* A *est un intervalle de* $\mathbb{R}$*, on note* $\mathbb{I}_A$ *la fonction indicatrice de* A *définie par* $\forall y \notin A,\ \mathbb{I}_A(y) = 0$ *et* $\forall y \in A,\ \mathbb{I}_A(y) = 1$.

Indication pour la question 2 : on pourra remarquer que q_θ *peut s'écrire sous la forme :*

$$\forall y \in \mathbb{R},\ q_\theta(y) = g(y-\theta),\ \text{ où }\ g(z) = \frac{3}{4}\left(1 - z^2\right) \mathbb{I}_{[-1,1]}(z).$$

1. Représentez sommairement les fonctions g et q_2 sur un graphique.

Corrigé – Voir la figure 1.

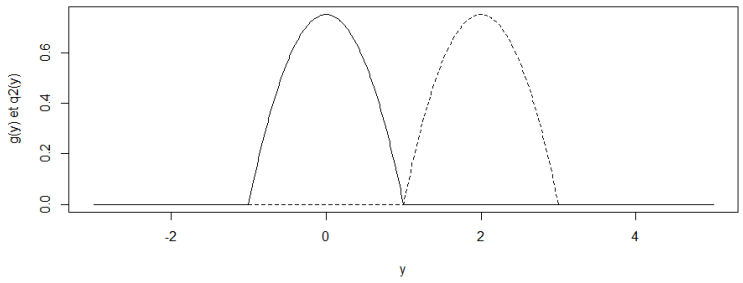

FIGURE 1: Fonctions g (ligne pleine) et q2 (tirets)

2. Vérifiez que q_θ est bien une densité, puis calculer les espérances : $E(X_1)$ et $E(X_1^2)$.

Corrigé – Il s'agit de calculer trois intégrales concernant q_θ, on se ramène à g par changement de variable $u = y - \theta$

$$\begin{aligned}\int_{\mathbb{R}} q_\theta(y)dy &= \int_{\mathbb{R}} g(u)du,\\ \int_{\mathbb{R}} yq_\theta(y)dy &= \int_{\mathbb{R}} (u+\theta)g(u)du = \theta\int_{\mathbb{R}} g(u)du + \int_{\mathbb{R}} ug(u)du,\\ \int_{\mathbb{R}} y^2q_\theta(y)dy &= \int_{\mathbb{R}} (u+\theta)^2g(u)du = \theta^2\int_{\mathbb{R}} g(u)du + 2\theta\int_{\mathbb{R}} ug(u)du + \int_{\mathbb{R}} u^2g(u)du.\end{aligned}$$

Des calculs simples donnent :

$$\int_{\mathbb{R}} g(y)dy = \int_{-1}^{1}\frac{3}{4}(1-y^2)dy = 1,\ \int_{\mathbb{R}} yg(y)dy = \int_{-1}^{1}\frac{3}{4}(y-y^3)dy = 0,\ \int_{\mathbb{R}} y^2g(y)dy = \int_{-1}^{1}\frac{3}{4}(y^2-y^4)dy = \frac{1}{5}.$$

Ainsi q_θ est une densité (fonction positive d'intégrale 1), $E(X_1) = \theta$, et $E(X_1^2) = \theta^2 + \frac{1}{5}$.

3. Montrer que la méthode des moments nous permet de construire deux estimateurs de θ, notés $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$, et donnés par :

$$\hat{\theta}^{(1)} := \frac{1}{9}\sum_{i=1}^{9} X_i, \quad \hat{\theta}^{(2)} := \sqrt{\frac{1}{9}\sum_{i=1}^{9} X_i^2 - \frac{1}{5}}.$$

Corrigé – On inverse la relation liant respectivement $E(X_1)$ puis $E(X_1^2)$) avec θ. C'est possible ici (on rappelle qu'on a supposé $\theta \in \mathbb{R}_+^*$).

On obtient respectivement $\theta = E(X_1)$ et $\theta = \sqrt{E(X_1^2) - \frac{1}{5}}$.

On définit l'estimateur en remplaçant respectivement $E(X_1)$ puis $E(X_1^2)$ par $\bar{X} = \frac{\sum_{i=1}^{9} X_i}{9}$ puis par $\frac{\sum_{i=1}^{9} X_i^2}{9}$, d'où les deux estimateurs demandés.

4. On admettra dans cette question que

$$\forall t > 0,\ \forall y_1, \dots, y_9 \in [\theta - 1, \theta + 1],\ \prod_{i=1}^{9} \mathbb{I}_{[t-1,t+1]}(y_i) = \mathbb{I}_{[-1+\max_{i=1,\dots,9} y_i, 1+\min_{i=1,\dots,9} y_i]}(t).$$

Donner l'expression de la fonction de vraisemblance associée aux données x.

Corrigé – Pour trouver la fonction de vraisemblance, il suffit d'écrire la densité du vecteur X :

$$\forall y_1, \dots, y_9 \in \mathbb{R},\ f_X(y_1, \dots, y_9) = \prod_{i=1}^{9} q_\theta(y_i),$$

puis on change les arguments : $y_1, \dots, y_9$ sont remplacées par les données $x_1, \dots, x_9$ et θ est remplacée par une variable muette $t \in \mathbb{R}_+^*$. On obtient :

$$\begin{aligned}\forall t \in \mathbb{R}_+^*,\ L(t) = \prod_{i=1}^{9} q_t(x_i) &= \frac{3^9}{4^9}\prod_{i=1}^{9}(1-(t-x_i)^2)\mathbb{I}_{\left[-1+\max_{i=1,\dots,9} x_i, 1+\min_{i=1,\dots,9} x_i\right]}(t)\\ &= \frac{3^9}{4^9}[1-(t-4.52)^2][1-(t-5.72)^2]\dots[1-(t-5.8)^2]\mathbb{I}_{[4.8,5.5]}(t).\end{aligned}$$

Pour information l'allure de la fonction est donnée dans la figure 2.

Cette fonction est difficile à étudier de façon théorique, mais on peut la représenter à l'aide d'un logiciel. On peut calculer aussi par logiciel que cette fonction est maximale environ en l'abscisse $t = 5.8$.

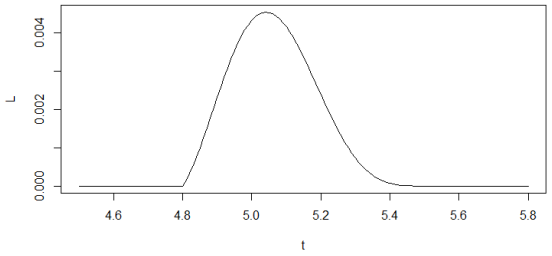

FIGURE 2: Fonction de vraisemblance L(t) pour t variant de 4.5 à 5.8

Corrigé – ATTENTION ERRATUM : il y a eu une erreur de programmation sur la fonction de vraisemblance ! En réalité la fonction est maximale environ en l'abscisse $t = 5.04$.

5. Donner les trois estimations numériques, à partir des données x, fournies par les deux méthodes des moments, et par la méthode de maximum de vraisemblance.

Corrigé – Application numérique : les trois estimations sont respectivement $\frac{1}{9}\sum_{i=1}^{9} x_i = 4.95$, $\sqrt{\frac{1}{9}\sum_{i=1}^{9} x_i^2 - \frac{1}{5}} = 4.95$ *et* 5.8.

ATTENTION ERRATUM : en réalité 5.04 *pour le 3ème.*

6. Dans cette question seulement, on suppose qu'on peut réitérer les expériences un nombre n arbitraire de fois : on remplace 9 par n dans les expressions de $\hat{\theta}^{(1)}$ et de $\hat{\theta}^{(2)}$. Montrer que $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$ sont consistants.

Corrigé – Cela découle de la loi des grands nombres (vérifier bien que les hypothèses sont satisfaites, voir cours et TD) appliquée respectivement aux variables X_i *et* $Y_i := X_i^2$, *et du fait que la fonction* $u \to \sqrt{u^2 - \frac{1}{5}}$ *est continue.*

7. Déterminer le biais, la variance et le risque quadratique (ou erreur quadratique moyenne) de $\hat{\theta}^{(1)}$.

Corrigé – Biais : $B(\hat{\theta}^{(1)}) = E(\hat{\theta}^{(1)}) - \theta = \frac{\sum_{i=1}^{9} E(X_i)}{9} - \theta = \theta - \theta = 0$ *par linéarité de l'espérance et vu que* $E(X_i) = \theta$.

Variance : $Var(\hat{\theta}^{(1)}) = \frac{Var(\sum_{i=1}^{9} X_i)}{81} = \frac{\sum_{i=1}^{9} Var(X_i)}{81} = \frac{9/5}{81} = 0.022$ *car les variables dans la somme sont indépendantes.*

Risque quadratique : $RQ(\hat{\theta}^{(1)}) = B(\hat{\theta}^{(1)})^2 + Var(\hat{\theta}^{(1)}) = 0.022$ *d'après la décomposition biais - variance.*

8. Le calculs théoriques du biais et du risque quadratique des deux autres estimateurs sont impossibles, mais on peut en trouver des approximations par des simulations sur ordinateur. On obtient les résultats suivants :

Type d'estimateur	Biais	Risque quadratique
moment d'ordre 2	-0.00083	0.02198
maximum de vraisemblance	0.02098	0.20187

Commentez : lequel des trois estimateurs vous semble préférable ?

Corrigé – Les deux estimateurs par les méthode des moments ont des qualités comparables (biais proches de 0, *variance de l'ordre de 0.02). L'estimateur par la méthode de max de vraisemblance est catastrophique : biais et variance beaucoup plus fortes, et en plus il est difficile à calculer.*
On donne donc plus de crédit à l'estimation 1 ou 2 : $\theta \approx 4.95$.

ATTENTION ERRATUM : il y a eu une erreur de programmation sur la fonction de vraisemblance ! En réalité la qualité estimée de l'estimateur par vraisemblance est donnée par :

Type d'estimateur	*Biais*	*Risque quadratique*
maximum de vraisemblance	*-0.0011*	*0.01757*

Donc il est légèrement meilleur que les deux autres estimateurs, au niveau du risque quadratique.