Contrôle des connaissances du cours « Méthodes de traitement des données » du DEA de Strasbourg.

Corrigé

Le jeudi 28 janvier 1999 de 10h 30à 12h30 à l'Observatoire de Strasbourg.

Le contrôle est noté sur 20, le barème est indiqué à la fin de chaque groupe de questions. Les exercices sont indépendants. Le polycopié du cours et les notes de cours sont les seuls documents autorisés.

Simulation de la loi normale.

Le but de cet exercice est de trouver un moyen de simuler une variable aléatoire suivant une loi normale réduite (plus exactement centrée et réduite) lorsque l'on dispose d'un générateur de nombres aléatoires suivant la loi uniforme entre 0 et 1.

On rappelle qu'une variable aléatoire X suit la loi normale réduite si X suit une loi normale de moyenne 0 et d'écart type 1.

① Soit X une variable aléatoire suivant la loi normale réduite, quelle est sa densité de probabilité $f_x(x)$?

② Soit (X, Y) un couple de variables aléatoires indépendantes, quelle est la densité de probabilité $f(x, y)$ de ce couple ?

③ On désire exprimer les coordonnées de ce couple, non pas suivant les coordonnées rectangulaires mais suivant les coordonnées polaires définies par le changement de variable :

$$\begin{cases} x = r\cos\theta\,, \\ y = r\sin\theta\,. \end{cases}$$

Quelle est la densité de probabilité $g(r, \theta)$ du couple de variables aléatoires (R, Θ) définies par : $X = R\cos\Theta$ et $Y = R\sin\Theta$?

④ Montrez que les variables aléatoires R et Θ sont indépendantes et donnez leur densité de probabilité ?

⑤ On désire simuler le couple (X, Y) en passant par les coordonnées polaires. On dispose d'un générateur de nombres aléatoires U suivant la loi uniforme entre 0 et 1. Soit u un nombre issu de ce générateur (ou variable) aléatoire U.

Comment transformer ces nombres u afin que les nombres θ suivent la loi de l'argument Θ ?

⑥ Même question mais pour simuler R ?

⑦ Comment peut-on simuler le couple (X, Y) à l'aide du couple (R, Θ) ?

⑧ Admettons que l'on dispose d'un moyen très rapide de simuler un couple (X', Y') de variables aléatoires indépendantes tel que le point de coordonnées (X', Y') soit réparti uniformément sur le disque unité : $X'2 + Y'2 \leq 1$.

Comment tirer parti de la connaissance du couple (X', Y') de façon à simplifier et accélérer le calcul évoqué à la question précédente ?

$(1+1+1+1+1+2+2+1=10$ points.$)$

Corrigé.

❶ C'est une simple question de cours :

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}x^2\}\,.$$

❷ La densité de probabilité d'un couple est le produit de la densité de probabilité des lois marginales de chacun des termes de ce couple. Ici ces lois marginales sont des lois normales réduites, il vient :

$$\begin{aligned} f(x,y) &= \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}x^2\} \frac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2}y^2\}\,, \\ &= \frac{1}{2\pi} \exp\{-\tfrac{1}{2}(x^2+y^2)\}\,. \end{aligned}$$

❸ On peut utiliser la formule générale du cours ou bien raisonner comme suit. Le changement de variable est une bijection sauf en 0 qui est de mesure nulle, on peut alors écrire : $g(r,\theta)drd\theta = f(x,y)dxdy$ qui exprime la conservation de la probabilité d'un certain événement auquel sont attachées les variables x et y ou r et θ. On sait que $dxdy$, qui est l'élément de surface, vaut $rdrd\theta$ en polaires d'où : $g(r,\theta)drd\theta = f(x,y)rdrd\theta$. On en tire la densité demandée :

$$\begin{aligned} g(r,\theta) &= rf(x,y)\,, \\ &= rf(r\cos\theta, r\sin\theta)\,. \end{aligned}$$

Notez bien qu'en procédant ainsi il peut apparaître un problème de signe, celui-ci est levé en remarquant qu'une probabilité est toujours positive.

Finalement la densité cherchée est :

$$g(r,\theta) = \frac{1}{2\pi} r \exp\{-\tfrac{1}{2}r^2\}\,.$$

❹ La densité de probabilité du couple peut se mettre sous la forme d'un produit de deux densités de probabilité, l'une ne dépendant que de r et l'autre que de θ ; les variables aléatoires R et Θ sont donc indépendantes. Leurs densités de probabilité sont :

$$\begin{aligned} g_r(r) &= r\exp\{-\tfrac{1}{2}r^2\}\,, \\ g_\theta(\theta) &= \frac{1}{2\pi}\,. \end{aligned}$$

La seule difficulté est de faire en sorte que ces densités de probabilité soient normalisées. C'est évident dans le cas de g_θ et par conséquent g_r est aussi normalisée, il est inutile de le vérifier en calculant l'intégrale $\int_0^\infty r\exp\{\frac{1}{2}r^2\}\,dr$.

Notez que Θ, l'argument du couple de variables aléatoires (X,Y), suit la loi uniforme entre 0 et 2π.

❺ Il faut transformer une variable aléatoire uniforme entre 0 et 1, en une variable aléatoire uniforme entre 0 et 2π, le procédé est évident. Pour chaque u on fabrique un θ suivant la formule : $\theta = 2\pi u$.

❻ L'opération est plus délicate, il faut se souvenir que si X est une variable aléatoire continue de fonction de répartition F alors la variable aléatoire $U = F(X)$ est uniforme entre 0 et 1. Réciproquement si U est uniforme entre 0 et 1, $X = F^{-1}(U)$ suit la loi de fonction de répartition F.

Il vient :

$$\begin{aligned} u = F(r) &= \int_0^r t \exp\{-\tfrac{1}{2}t^2\}\, dt\,, \\ &= \int_0^{\frac{1}{2}r^2} \exp\{-v\}\, dv\,, \\ &= 1 - \exp\{-\tfrac{1}{2}r^2\}\,. \end{aligned}$$

D'où l'on tire la simulation demandée :

$$r = \sqrt{-2\ln(1-u)}\,.$$

Notez que si U est uniforme, $1-U$ l'est aussi par conséquent la formule précédente peut s'écrire plus simplement :

$$r = \sqrt{-2\ln u}\,.$$

❼ Le générateur de nombres aléatoires U est tel que la suite des réalisations : $u_1, u_2, u_3, u_4, \ldots$ peuvent être considérée comme des réalisations de variables aléatoires uniformes indépendantes. Désignons par U_1 la variable engendrant la suite des nombres de rang impairs et par U_2 celle correspondant aux rangs pairs. D'après le théorème de Slutski sur le changement de variable, les variables $\Theta = 2\pi U_1$ et $R = \sqrt{-2\ln U_2}$ sont aussi indépendantes. (Ce fait est quasi-évident et on ne vous en voudra pas si vous n'avez pas fait explicitement référence au théorème de Slutski.) Le couple (X, Y) est simulé grâce au changement de variables :

$$\begin{cases} X = \sqrt{-2\ln U_2}\cos(2\pi U_1)\,, \\ Y = \sqrt{-2\ln U_2}\sin(2\pi U_1)\,. \end{cases}$$

On retrouve ainsi une formule donnée dans le cours.

❽ Par symétrie l'argument Θ' du couple (X', Y') est réparti uniformément entre 0 et 2π, on peut alors poser : $\cos(2\pi U_1) = X'$ et $\sin(2\pi U_1) = Y'$, d'où on en déduit une façon rapide de simuler un couple de variables aléatoires indépendantes suivant la loi normale réduite :

$$\begin{cases} R = \sqrt{-2\ln U}\,, \\ X = R\,X'\,, \\ Y = R\,Y'\,. \end{cases}$$

Ce qui évite de calculer les fonctions trigonométriques. Pour remplir uniformément un disque unité on peut procéder ainsi : on tire deux variables aléatoires uniformes U et V entre -1 et 1 et on ne retient que les couples pour lesquels $U^2 + V^2 \leq 1$. Pour ceux-là on pose $X' = U$ et $Y' = V$.

Un modèle à 2 paramètres.

Un expérimentateur a réalisé l'ajustement de données $y_i,\ i = 1, \ldots, n$ à l'aide d'un modèle μ_i dépendant de 2 paramètres θ_1 et θ_2.

① L'expérimentateur a calculé la matrice $\boldsymbol{V}$ des variances-covariances des paramètres θ et il a obtenu la valeur suivante :

$$\boldsymbol{V} = \begin{pmatrix} 4 & 7 \\ 7 & 9 \end{pmatrix} .$$

Pouvez-vous dire s'il est possible que cette matrice soit bien une matrice des variances-covariances.

② Pour un autre jeu de données, l'expérimentateur a obtenu une estimation des θ à l'aide d'une méthode de moindres-carrés linéaire. Cette méthode lui a donné les estimations : $\hat{\theta}_1 = 1$, $\hat{\theta}_2 = 2$, cette fois-ci la matrice des variances-covariances vaut :

$$\boldsymbol{V} = \begin{pmatrix} 4 & 4 \\ 4 & 9 \end{pmatrix} .$$

Des considérations théoriques favorisent l'hypothèse que $\theta_1 = \theta_2 = 1$. Les résultats obtenus, ainsi que l'ajustement qui a été fait sont-ils en contradiction avec l'hypothèse théorique au niveau 90% de confiance ?

($1 + 3$ points.)

Corrigé.

❶ Une matrice des variances-covariances à deux paramètres s'écrit toujours :

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} ,$$

avec $|\rho| \leq 1$. Avec les données du problème on calcule : $\rho = 7/(2 \times 3) = 7/6$ qui est plus grand que 1. Cette matrice ne saurait être une matrice des variances-covariances, l'expérimentateur a commis une erreur dans l'évaluation des éléments de $\boldsymbol{V}$.

Notez qu'il y a d'autres moyens pour voir que $\boldsymbol{V}$ ne convient pas. Il suffit de démontrer qu'elle n'est pas définie positive, par exemple son déterminant vaut -13 qui n'est pas positif, ou encore la forme quadratique $\boldsymbol{x}^T\boldsymbol{V}\boldsymbol{x}$ avec $\boldsymbol{x}^T = (1 \ -1)$ n'est pas positive (elle vaut -1).

❷ Il faut calculer le χ^2, c'est-à-dire : $\chi^2 = \boldsymbol{u}^T\boldsymbol{V}^{-1}\boldsymbol{u}$ et le comparer au χ^2_γ attendu au niveau de confiance γ. On a $\chi^2_\gamma = F^{-1}(\gamma)$ où F est la fonction de répartition d'une variable aléatoire suivant une loi du χ^2 à 2 degrés de liberté.

Le vecteur $\boldsymbol{u} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ est l'écart entre le modèle théorique et ce que donne l'ajustement et $\boldsymbol{V}^{-1}$ est l'inverse de la matrice des variances-covariances, on a :

$$\boldsymbol{u} = \begin{pmatrix} 1-1 \\ 1-2 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} , \quad \boldsymbol{V}^{-1} = \begin{pmatrix} 0.45 & -0.2 \\ -0.2 & 0.2 \end{pmatrix} .$$

Avec ces valeurs on obtient $\chi^2 = 0.2$, la valeur de $\chi^2_{0.9}$ est donnée dans le polycopié à la fin du chapitre sur la loi normale, on y trouve que $\chi^2_{0.9} \approx (2.146)^2$. On est dans le cas où $\chi^2 < \chi^2_{0.9}$, l'hypothèse que $\theta_1 = \theta_2 = 1$ est compatible avec les données (sous réserve que le modèle linéaire est correct.)

Recherche de raies dans un spectre gamma.

Dans le domaine gamma, la détection d'un photon s'accompagne d'une mesure de son énergie avec une précision qui dépend du système de détection utilisé et qui varie de 0.1 % à 10 %. Le spectre d'une source dans le domaine 375 - 425 keV est le suivant :

E	386	387	388	389	390	391	392	393	394	395
N	115	118	95	90	100	104	93	97	104	94
E	396	397	398	399	400	401	402	403	404	405
N	130	140	93	106	107	73	109	102	97	99
E	406	407	408	409	410	411	412	413	414	415
N	108	112	90	125	114	84	89	97	107	103

ou E est l'énergie en keV et N le nombre de photons détectés entre $E - 0.5$ keV et $E + 0.5$ keV. Le signal détecté est la superposition de la source dont on mesure le spectre, et de bruit. Dans les deux cas, on suppose qu'il n'y a pas de corrélation dans le temps d'arrivée des photons, c'est à dire que la probabilité d'observer un photon dans l'intervalle de temps $[t_0, t_0 + \Delta t]$ ne dépend que de Δt.

① Quelle est la loi de probabilité suivie par N dans chacun des canaux d'énergie ?

② Peut-on donner une forme approchée de cette loi au vu des nombres de photons détectés ?

③ Il semble exister un excès à 396 et 397 keV, qui pourrait correspondre à la raie d'annihilation à 511 keV décalée vers le rouge par un effet relativiste ; la mesure de ce dernier permettant de mettre une contrainte sur la masse et le rayon de l'objet émetteur. On se propose de vérifier si cet excès est statistiquement significatif. On supposera que, en l'absence de tout signal, le nombre moyen de photons détectés dans chaque canal d'énergie est indépendant de E. Peut-on estimer la probabilité d'avoir obtenu 270 photons dans les canaux a 396 et 397 keV en l'absence de signal ?

④ On peut aussi estimer la probabilité que le signal obtenu résulte de la loi de probabilité déterminée en (1) avec une moyenne indépendante de l'énergie. Quelle est-elle ?

⑤ Les réponses aux questions (1) et (2) sont très différentes. Pouvez vous expliquer pourquoi ?

⑥ De quelle information sur la mesure de l'énergie de chaque photon détecté souhaiteriez-vous disposer pour conclure sur la réalité du signal ?

(1+1+1+1+1+1=6 points.)

Corrigé.

❶ N suit une loi de Poisson dans chaque canal d'énergie.

❷ N étant grand, la loi de Poisson peut être approchée par une loi de Gauss dont la moyenne est celle de la loi de Poisson et l'écart type la racine carrée de la moyenne.

❸ Une estimation de la moyenne de la loi de Poisson est

$$\mu = 1/30 \Sigma_{i=386}^{i=415} N_i = 103.1667$$

Le nombre de photons attendus dans les canaux 396 et 397 est donc de 206.33, et suit une loi de Poisson, que l'on approchera par une gaussienne. On se situe à 4.43 écarts types de la moyenne, ce qui correspond à une probabilité $P_1 = 9.4 \times 10^{-6}$.

❹ Il faut effectuer le test du χ^2, donné par

$$\chi^2 = 1/30 \Sigma_{i=386}^{i=415} (N_i - \mu)^2 / \mu = 53.60$$

Il y a 29 degrés de liberté. Ce nombre étant grand, on peut approximer la loi du χ^2 par une gaussienne dont la moyenne est 29 et l'écart type $\sqrt{58}$; la déviation à la moyenne est de 3.2 σ, ce qui correspond à une probabilité $P_2 = 1.4 \times 10^{-3}$.

❺ La différence sensible tient au fait d'une part que dans le premier cas on spécifie la position des canaux où existe un excès, qui n'est en principe pas connu. La probabilité d'existence d'un signal est donc de l'ordre de $30P_1$. D'un autre côté, le test du χ^2 ne tient pas compte du fait que les deux excès les plus importants sont consécutifs, et sous-estime donc la probabilité d'existence d'un signal.

❻ On souhaiterait connaître la résolution spectrale du détecteur. La présence d'une raie ne peut être envisagée que si celle-ci est au plus de l'ordre du keV.