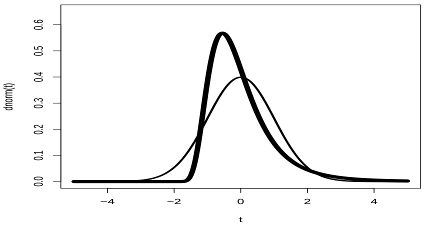
Examen: Modélisation aléatoire, option statistiques

Ecole Centrale

24 Septembre 2008

1 Moments



Le graphe le plus fin représente la densité de la loi normale centrée réduite. Le plus épais représente la densité d'une loi d'espérance 0 et de variance 1. Que peut-on dire de sa torsion (skewness) et de sa kurtosis ? Quel type de procédure mettriez-vous en oeuvre pour estimer la moyenne d'un échantillon dont la loi serait semblable à celle-ci ?

2 Estimer une moyenne et une variance

1. Montrer que si $X_1, \ldots, X_n$ est un échantillon de loi $\mathcal{N}(0,1)$, alors
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(0, 1/n), \text{ et}$$
$$(n-1)\hat{\sigma}_n^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \sim \chi^2(n-1).$$
Montrer qu'en outre $\bar{X}_n$ et $\hat{\sigma}_n^2$ sont indépendantes.
2. On ne suppose plus que les variables X_i sont gaussiennes, mais elles sont toujours indépendantes de moyenne 0 et de variance 1. Calculer l'espérance de $\bar{X}_n$ et de $\hat{\sigma}_n^2$, la variance de $\bar{X}_n$ ainsi que la covariance $Cov(\bar{X}_n, \hat{\sigma}_n^2)$. Les variables aléatoires $\bar{X}_n$ et $\hat{\sigma}_n^2$ sont-elles indépendantes ?

3 Régression jointe

Des études ont montré que la durée de vie d'un rat exposé 2h au soleil chaque jour est une variable aléatoire de loi normale $\mathcal{N}(\mu, \sigma^2)$. On pense qu'une exposition plus longue au soleil augmente la durée de vie des rats, et qu'inversement une exposition moins longue a tendance à la diminuer. Pour modéliser simplement la situation, on considère donc que la durée de vie Y_i du rat numéro i est une variable aléatoire de loi $\mathcal{N}(\mu + \alpha(x_i - 120), \sigma^2)$, où x_i est la durée (en minutes) d'exposition du rat au soleil chaque jour et où α est un facteur réel à estimer.

1. Pour étudier cette dépendance, on constitue un échantillon de $n = 59$ rats, dont on expose le rat numéro $i \in \{1, \ldots, n\}$ à $x_i = 90 + i$ minutes de soleil. On relève les résultats $Y_1, \ldots, Y_n$. Proposer des estimateurs $\hat{\mu}$, $\hat{\alpha}$ et $\hat{\sigma}^2$ de μ, α et σ^2, préciser quelle est leur loi, et expliquer comment tester l'absence d'influence de la durée d'exposition au soleil sur la durée de vie des rats.
2. Dans un deuxième temps, on recommence la même étude sur $n = 59$ autres rats à la ration alimentaire desquels on ajoute une molécule censée modifier la production de mélanine. En présence de cette molécule, on suppose que la durée de vie Y'_k d'un rat suit cette fois une loi $\mathcal{N}(\mu + \beta(x_i - 120), \sigma^2)$, où β est un facteur réel à estimer. Expliquer comment construire, *à partir des deux échantillons* $(Y_k)_{1 \le k \le n}$ *et* $(Y'_k)_{1 \le k \le n}$, des estimateurs $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$ et $\hat{\sigma}^2$ de μ, α, β et σ^2. Expliquer comment tester l'efficacité de la molécule, c'est-à-dire $\beta > \alpha$.

4 Statistiques bayésiennes

Les pièces fabriquées dans une usine sont inspectées par une machine. Chaque pièce a une probabilité $\theta \in [0,1]$ d'être défectueuse, et on considère que chaque pièce est défectueuse ou non indépendamment des autres. On note X_1 le nombre de pièces inspectées par la machine jusqu'à ce qu'elle trouve la première pièce défectueuse ; on note ensuite X_2 le nombre de pièces inspectées par la machine, à partir de la (X_1+1)-ième, jusqu'à ce qu'elle trouve la deuxième pièce défectueuse, etc...

1. Montrer que les variables aléatoires X_i sont indépendantes et de même loi. Préciser quelle est cette loi et calculer son espérance et sa variance. Montrer que la densité des observations $X_1 = x_1, \dots, x_n = x_n$ est :
$$p(x_1, \dots, x_n|\theta) = \theta^n (1-\theta)^{s_n - n}, \quad \text{avec } s_n = \sum_{i=1}^{n} x_i.$$
2. Dans cette question et dans les suivantes, on veut faire une estimation bayésienne de p. L'idée qu'on se fait avant toute expérience du paramètre θ nous incite à choisir comme loi a priori (prior) la loi Beta$(2, 100)$, dont la densité par rapport à la mesure de Lebesgue est :
$$\pi(\theta) = \frac{\theta(1-\theta)^{99}}{\beta(2,100)} \mathbb{1}_{[0,1]}(\theta).$$
Tracer l'allure de la fonction π. A combien estime-t-on *a priori* la probabilité qu'une machine soit défectueuse ?
3. Calculer la densité jointe du paramètre θ et des observations $X_1 = x_1, \dots, X_n = x_n$, puis montrer que la loi de θ a posteriori (sachant $X_1 = x_1, \dots, X_n = x_n$) est la loi Beta$(2+n, 10+s_n-n)$, avec $s_n = \sum_{i=1}^n x_i$.
4. En déduire l'estimateur un estimateur a posteriori de θ. Le comparer à l'estimateur du maximum de vraisemblance.

5 Débruitage en base de Fourier

On cherche à reconstruire une fonction 1-périodique $f : [0,1] \to \mathbb{R}$ continue à partir d'un échantillonage bruité
$$Y_k = f\left(\frac{k}{n}\right) + \sigma Z_k, \quad 0 \le k \le n-1,$$
où n est un entier strictement positif pair, σ un réel strictement positif, et $(Z_k)_{0 \le k \le n-1}$ est un n-échantillon de variables aléatoires gaussiennes centrées réduites. Pour $k \in \mathbb{N}$, on pose
$$\tilde{c}_k = \begin{pmatrix} \cos\left(\frac{0 \times 2\pi k}{n}\right) \\ \cos\left(\frac{1 \times 2\pi k}{n}\right) \\ \cos\left(\frac{2 \times 2\pi k}{n}\right) \\ \vdots \\ \cos\left(\frac{(n-1) \times 2\pi k}{n}\right) \end{pmatrix} \in \mathbb{R}^n \quad \text{et} \quad \tilde{s}_k = \begin{pmatrix} \sin\left(\frac{0 \times 2\pi k}{n}\right) \\ \sin\left(\frac{1 \times 2\pi k}{n}\right) \\ \sin\left(\frac{2 \times 2\pi k}{n}\right) \\ \vdots \\ \sin\left(\frac{(n-1) \times 2\pi k}{n}\right) \end{pmatrix} \in \mathbb{R}^n$$
On pose également $c_k = \tilde{c}_k / \|\tilde{c}_k\|$ et $s_k = \tilde{s}_k / \|\tilde{s}_k\|$, où $\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}$ désigne la norme euclidienne d'un vecteur $x \in \mathbb{R}^n$. Soit enfin
$$\Omega_n = \{c_k : 0 \le k \le n/2\} \cup \{s_k : 1 \le k \le n/2 - 1\}.$$

1. Expliciter Ω_2 et Ω_4.
2. Pour $k, l \in \{0, \dots, n/2\}$, calculer les produits scalaires $\langle \tilde{c}_k, \tilde{c}_l \rangle$, $\langle \tilde{s}_k, \tilde{s}_l \rangle$ et $\langle \tilde{c}_k, \tilde{s}_l \rangle$.
3. Montrer que pour tout entier n pair (non nul), Ω_n est une base orthonormée de $\mathbb{R}^n$.
4. Décrire une procédure qui prend en entrée un vecteur $Y = (Y_k)_{0 \le k \le n-1}$ et qui renvoie un vecteur d'estimateurs $\left(\hat{f}\left(\frac{0}{n}\right), \hat{f}\left(\frac{1}{n}\right), \dots, \hat{f}\left(\frac{n-1}{n}\right)\right)$.