UNIVERSITÉ DE NANTES

Préparation à l'Agrégation

TP Modélisation Statistique

P. Rochet

Quelques commandes Scilab

Génération de nombres aléatoires

La fonction `grand(m,n,'*',p1,p2,...)` sert à générer une matrice de taille $m \times n$ dont les entrées sont des réalisations indépendantes de loi * de paramètres $p1, p2, ...$ (ex: `def` = loi uniforme (on peut utiliser également la commande `rand`), `nor` = loi normale, `'mn'` = vecteur Gaussien, `exp` = loi exponentielle, `bin` = loi binomiale, etc...).

Représentation graphique

Pour représenter graphiquement une loi discrète, on peut utiliser la commande `plot2d3` (diagramme en bâton) ou `bar` (diagramme en barres).

Pour représenter graphiquement une loi continue, on peut utiliser la commande `plot2d`.

Pour construire un histogramme, on peut utiliser la commande `histplot` en précisant en premier argument le nombre de classes.

Fonctions de répartition

Les fonctions de répartition des lois usuelles sont préprogrammées dans Scilab sont le nom `cdf*` (ex: `cdfnor`, `cdfbet`...).

Le premier argument en entrée est la chaîne `'PQ'` si on veut évaluer la fonction de répartition elle-même. Par exemple la commande `x=linspace(-3,3); y=cdfnor('PQ',x,zeros(x),ones(x))` renvoie la fonction de répartition de la loi normale centrée réduite sur l'intervalle $(-3, 3)$. Les densités des lois usuelles ne sont pas préprogrammées.

La commande `cdf*` permet également d'inverser la fonction de répartition et donc de déterminer les quantiles des lois usuelles. On donne alors comme premier argument la chaîne de caractères `'X'`. Par exemple la commande `cdfnor('X',0,1,0.975,0.025)` renvoie le quantile à 95% de la loi normale centrée réduite.

Chaînes de Markov

La commande `grand(n,'markov',P,x)` où P est une matrice stochastique d'ordre N et $x \in \{1, ..., N\}^m$ permet de générer m trajectoires de taille n d'une chaîne de Markov à N états de matrice de transition P et d'états initiaux $x_1, ..., x_m$.

La commande `genmarkov([n1,n2,...,np],nt)` génère une matrice de transition aléatoire d'une chaîne de Markov ayant p classes récurrentes de tailles respectives $n_1, ..., n_p$ et n_t états transients communiquant tous entre eux.

1 Simulation de variables aléatoires

Exercice 1: Méthode d'inversion. Soit X une variable aléatoire de fonction de répartition F. On note F^- son inverse généralisée

$$F^-(u) = \inf\{x \in \mathbb{R} | F(x) \geq u\}.$$

1. Soit $U \sim \mathcal{U}[0,1]$, montrer que $F^-(U)$ est une variable aléatoire de fonction de répartition F.
2. Si F est continue, montrer que $F(X)$ est une variable aléatoire uniforme sur $[0,1]$.
3. A partir de la commande `rand` qui génère des variables aléatoires de loi uniforme, simuler un échantillon de taille $n = 500$ de loi exponentielle de paramètre $\lambda = 2$.

Exercice 2: Méthode de rejet. Soient f et g deux densités de probabilité, M un réel positif tel que $f \leq Mg$, $(Y_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes de densité g et $(U_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$ (et indépendantes des Y_i).

1. Calculer la loi de
$$T = \inf\left\{n \geq 1 | U_n \leq \frac{f(Y_n)}{Mg(Y_n)}\right\}$$
et montrer que $X := Y_T$ est une va de densité f.
2. A partir d'un générateur de loi uniforme, simuler un échantillon iid de taille $n = 500$ de variables aléatoires de loi $\mathcal{N}(0,1)$ par la méthode de rejet, en remarquant par exemple que la densité Gaussienne standard est bornée par deux fois la densité de Cauchy.
3. Simuler un échantillon iid de taille $n = 200$ de loi uniforme sur la boule unité de $\mathbb{R}^2$.

Exercice 3: Mélange de lois. Soient Y_1 et Y_2 deux variables aléatoires réelles de densités respectives f_1 et f_2.

1. Soit X une variable de Bernoulli de paramètre p indépendante de Y_1 et Y_2, déterminer la densité de la variable aléatoire
$$Y = Y_1 X + Y_2(1 - X).$$
2. On note ϕ_{μ,σ^2} la densité d'une va de loi normale $\mathcal{N}(\mu, \sigma^2)$. Simuler un échantillon iid de taille $n = 1000$ de variables aléatoires de densité
$$f(x) = \frac{1}{3}\phi_{-2,1}(x) + \frac{2}{3}\phi_{3,4}(x)\ ,\ x \in \mathbb{R}.$$

Exercice 4: Méthode de Box-Muller. Soit (X, Y) un vecteur Gaussian standard de $\mathbb{R}^2$, càd un vecteur aléatoire de densité

$$f(x,y) = \frac{1}{2\pi}\exp(-\tfrac{x^2+y^2}{2}),\ x, y \in \mathbb{R}.$$

On considère les coordonnées polaires du vecteur (X, Y) en posant $X = R\cos(T)$ et $Y = R\sin(T)$ avec $R \in [0,1]$ et $T \in [0, 2\pi)$.

1. Montrer que R et T sont indépendantes et préciser leurs lois.
2. Soit (U, V) un couple de variables aléatoires indépendantes uniformes sur $[0,1]$. Déduire de la question précédente que le couple
$$(X, Y) = \left(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\right),$$
est un vecteur Gaussien standard de $\mathbb{R}^2$.

2 Méthodes de Monte-Carlo

Exercice 1: Loi des grands nombres. Simuler un échantillon de taille $n = 1000$ de va uniformes sur $[0, 2]$.

1. A l'aide de la fonction `cumsum()`, calculer les moyennes des k premières valeurs de l'échantillon pour k allant de 1 à 1000 et représenter graphiquement les valeurs obtenues.
2. Répéter la procédure 5 fois et superposer les graphiques.
3. Faire de même pour des échantillons de loi de Cauchy. Commenter.

Exercice 3: Théorème central limite. Simuler $N = 1000$ échantillons de taille $n = 50$ de va indépendantes de loi exponentielle de paramètre 2. Illustrer graphiquement le TCL en superposant l'histogramme des moyennes empiriques (renormalisées) obtenues pour chaque échantillon à la densité limite théorique.

Exercice 4: Théorème de Glivenko-Cantelli. Simuler un échantillon $X_1, ..., X_n$ de taille $n = 200$ de va indépendantes de loi normale $\mathcal{N}(0, 1)$.

1. Représenter la fonction de répartition empirique F_n de l'échantillon et la superposer avec la vraie fonction de répartition F.
2. Pour k allant de 1 à n, calculer la variable aléatoire

$$Y_k = \|F_k - F\|_\infty = \sup_{x \in \mathbb{R}} |F_k(x) - F(x)|,$$

   obtenue à partir des k premières valeurs de l'échantillons. Commenter.

Exercice 5: Théorème de Kolmogorov-Smirnov. Générer $N = 1000$ réalisations de la variable aléatoire Y_n définie dans l'exercice précédent, issue d'un échantillon de taille $n = 200$ de va indépendantes de loi normale $\mathcal{N}(0, 1)$.

1. Tracer l'histogramme.
2. Faire de même avec un échantillon de loi exponentielle puis de loi uniforme. Que constate-t-on?

Exercice 5: Calcul d'intégrale. Utiliser la méthode de Monte-Carlo pour calculer numériquement les intégrales suivantes

1. $I_1 = \displaystyle\int_{\mathbb{R}} e^{-2x^2} |\cos(x)| dx.$
2. $I_2 = \displaystyle\int_D e^{x^2 \cos(y)} \sqrt{1 + \log^2(y)}\, dxdy$, où $D = \{(x, y) \in [0, 1]^2 : x + \sin(y) \leq 1\}$.

3 Vecteurs Gaussiens

Exercice 1: Transformations affines de vecteurs Gaussiens. Soient $X_1, ..., X_k$ des va iid de loi $\mathcal{N}(0,1)$.

1. Montrer que $X = (X_1, ..., X_k)^\top$ est un vecteur Gaussien standard de $\mathbb{R}^k$: $X \sim \mathcal{N}(0, \mathrm{I})$.
2. Soit $\mathsf{B} \in \mathbb{R}^{p\times k}$ une matrice surjective et $\mathsf{m} \in \mathbb{R}^p$. Montrer que $Y = \mathsf{B}X + \mathsf{m} \sim \mathcal{N}(\mathsf{m}, \mathsf{B}\mathsf{B}^\top)$, c'est-à-dire, Y a pour densité
$$\frac{1}{\sqrt{2\pi}^p\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}(y-\mathsf{m})^\top\Sigma^{-1}(y-\mathsf{m})\Big)\ ,\ y \in \mathbb{R}^p,$$
où $\Sigma = \mathsf{B}\mathsf{B}^\top$. Que se passe-t-il si B n'est pas surjective?
3. En utilisant la commande `sqrtm` qui calcule la racine carrée d'une matrice symétrique définie positive, générer un vecteur Gaussien Y de paramètres
$$\mathsf{m} = \begin{bmatrix} 2 \\ 3 \end{bmatrix},\ \Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix},$$
à partir de variables aléatoires normales indépendantes.

Exercice 2: Loi du χ^2. La norme Eclidienne au carré d'un vecteur Gaussien standard de $\mathbb{R}^n$ suit une loi du χ^2 à n degrés de liberté, notée $\chi^2(n)$.

1. Montrer que la somme de deux variables aléatoires indépendantes de lois $\chi^2(n)$ et $\chi^2(n')$ respectivement, suit un $\chi^2(n+n')$.
2. Soit $Y \sim \mathcal{N}(\mathsf{m}, \Sigma)$ définie dans l'exercice précédent, trouver une fonction $f$telle que $f(Y) \sim \chi^2(2)$.

Exercice 3: Loi de Student. Soient X et Y deux variables aléatoires indépendantes de lois $\mathcal{N}(0,1)$ et $\chi^2(n)$ respectivement, la variable aléatoire
$$T = \frac{X}{\sqrt{Y/n}}$$
suit une loi de Student à n degrés de liberté, notée $\mathcal{T}(n)$.

1. Montrer qu'une suite de va T_n de lois respectives $\mathcal{T}(n)$, converge en loi vers une loi normale centrée réduite.
2. Ecrire une fonction qui à n associe une réalisation d'une variable aléatoire de loi $\mathcal{T}(n)$ construite à partir de variables Gaussiennes.
3. Superposer les densités de Student pour $n = 1, 5, 10$.

Exercice 4: Loi de Fisher. Soient X et Y deux va indépendantes de lois $\chi^2(n)$ et $\chi^2(n')$, la variable aléatoire
$$F = \frac{X/n}{Y/n'}$$
suit une loi de Fisher à n et n' degrés de liberté, notée $\mathcal{F}(n, n')$.

1. Ecrire une fonction qui à n, n' associe une réalisation d'une variable aléatoire de loi $\mathcal{F}(n, n')$ construite à partir de variables Gaussiennes.

Exercice 5: Echantillon Gaussien. Soit $X = (X_1, ..., X_n)^\top$ un échantillon iid de loi normale $\mathcal{N}(\mathsf{m}, \sigma^2)$ sur $\mathbb{R}$. On note $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

1. Soit $\mathbf{1} = (1, ..., 1)^\top$, montrer que les vecteurs $X - \overline{X}.\mathbf{1}$ et $\mathbf{1}$ sont orthogonaux dans $\mathbb{R}^n$.
2. En déduire que $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2$ est la norme au carré d'un vecteur Gaussien standard de $\mathbb{R}^{n-1}$.
3. Montrer que $X - \overline{X}.\mathbf{1}$ et $\overline{X}$ sont non-corrélés et donc indépendants puisque transformations linéaires d'un vecteur Gaussien.
4. En déduire la loi de
$$T = \sqrt{n}\frac{\overline{X} - \mathsf{m}}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2/(n-1)}}.$$
5. Vérifier graphiquement ce résultat pour $n = 10, \mathsf{m} = 1, \sigma^2 = 4$ par Monte-Carlo.

4 Chaînes de Markov

Exercice 1: Estimation de la probabilité de transition.

1. Générer aléatoirement la matrice de transition P d'une chaîne irréductible d'espace d'états à 4 éléments.
2. Calculer P^4 puis P^{10}. Commenter.
3. Générer une trajectoire de longueur $n = 10000$ d'une chaîne de Markov de matrice de transition P.
4. Quel est l'estimateur du maximum de vraisemblance $\hat{P}$ de P? Sous quelles conditions est-il convergent presque sûrement?
5. Calculer numériquement l'estimateur $\hat{P}$.

Exercice 2: Retournement du temps.

1. Générer aléatoirement la matrice de transition P d'une chaîne irréductible d'espace d'états à 5 éléments, puis générer une trajectoire de longueur $n = 10000$ de cette chaîne.
2. Construire la matrice $\hat{Q}$ définie par
$$\hat{Q}_{ij} = \frac{\sum_{k=2}^n \mathbb{1}\{X_k = i, X_{k-1} = j\}}{\sum_{k=2}^n \mathbb{1}\{X_k = i\}}.$$
3. Vérifier numériquement que $\hat{Q}_{ij}$ converge vers $Q_{ij} := P_{ji}\frac{\pi(j)}{\pi(i)}$ où π est la loi stationnaire de la chaîne initiale.
4. Quelle est la loi stationnaire de la chaîne de matrice de transition Q?

Exercice 3: Le modèle de Wright-Fisher. Une population de taille fixe 10 comprend deux types d'individus A et B. Le nombre d'individus A à un temps n est notée X_n. Chaque individu de la $(n+1)$-ième génération a un parent tiré au sort uniformément dans la population et il hérite du type de son unique parent.

1. Montrer que $(X_n)_{n\geq 1}$ est une chaîne de Markov dont on précisera l'espace d'états et la matrice de transition.
2. Quels sont les états absorbants? Montrer que les états non absorbants communiquent entre eux.
3. Décrire le comportement de la chaîne quand $n \to \infty$.
4. Simuler plusieurs trajectoires de cette chaîne avec état initial $m \in \{1, ..., 9\}$ et vérifier numériquement le résultat de la question 3.

5 Intervalles de confiance et tests

Exercice 1: Intervalle de confiance pour la moyenne. On observe un échantillon iid $X_1, ..., X_n$ de loi normale $\mathcal{N}(\mu, 2)$ où μ est inconnu.

1. Donner un intervalle de confiance à 95% pour μ.
2. Simuler 1000 échantillons de taille $n = 100$ en prenant $\mu = 1$ et vérifier numériquement la précision de l'intervalle de confiance.
3. Refaire l'expérience pour un échantillon de loi exponentielle de paramètre λ, en remarquant que la variable aléatoire $Y = \lambda \sum_{i=1}^{n} X_i$ suit une loi gamma $\Gamma(1, n)$.

Exercice 2: Intervalle de confiance d'une proportion. Soit $X_1, ..., X_n$ un échantillon iid de loi de Bernoulli $\mathcal{B}(p)$.

1. Construire un intervalle de confiance à 95% asymptotiquement pour p et vérifier par Monte-Carlo la précision de l'intervalle pour $n = 200$ et $p = 0.4$.
2. Représenter la longueur moyenne de l'intervalle de confiance lorsque $n = 200$, pour différentes valeurs de $p \in]0, 1[$.

Exercice 3: Text d'adéquation du χ^2. On souhaite vérifier la qualité du générateur de nombres aléatoires d'une calculatrice scientifique. Pour cela, on procède à 250 tirages dans l'ensemble $\{0, ..., 9\}$ et on obtient les résultats suivants:

x	0	1	2	3	4	5	6	7	8	9
$N(x)$	28	32	23	26	23	31	18	19	21	29

Tester si le générateur produit des entiers uniformément.

Exercice 4: Test de Kolmogorov-Smirnov.

1. Générer deux échantillons $X_1, ..., X_n$ et $Y_1, ..., Y_n$ de loi gamma de paramètres α et β.
2. Tester à l'aide d'un test de Kolmogorov-Smirnov si les variables $X_i/(X_i + Y_i)$ suivent une loi bêta $B(\alpha, \beta)$.

Aide. La fonction de répartition F_{ks} de la loi de Kolmogorov-Smirnov est codée par la commande `pks` de la `stixbox` dont l'installation se fait en tapant `atomsInstall("stixbox")` dans la console `scilab`. La p-value du test est alors donnée par $\text{pv} = 1 - F_{\text{ks}}(K_n)$ où K_n est la statistique de Kolmogorov-Smirnov.