Variable aléatoire et simulation : densité, histogramme, mélange, corrélation

Sujet de travail pratique

Le présent travail, pratique, concerne la simulation de variables aléatoires et poursuit un triple objectif.

1. D'une part, il vous familiarise avec la notion « expérimentale » de variables aléatoires (v.a.).
2. D'autre part, il vous initie à la programmation sous Matlab / Octave ou consolide vos compétences antérieures.
3. Naturellement, il prépare aussi la suite de l'enseignement consacré à l'estimation, notamment dans un cadre bayésien.

Nous allons voir comment générer des réalisations d'une variable aléatoire et comment l'histogramme des réalisations permet d'obtenir une approximation de la densité de probabilité. Enfin, la troisième partie abordera les notions de couple de v.a. et de corrélation. Deux densités de probabilité serviront d'illustration : la densité uniforme et la densité normale (ou gaussienne).

Préparation : avant la séance répondez aux questions 1, 2, 8a, 12a, 16, 17a et 18.

Dans un premier temps, on considère une v.a. X uniformément distribuée sur l'intervalle $[x_\mathrm{m}, x_\mathrm{M}]$. Sa densité de probabilité s'écrit comme suit.

$$f_X(x) = \begin{cases} \dfrac{1}{x_\mathrm{M} - x_\mathrm{m}} & \text{si } x \in [x_\mathrm{m}, x_\mathrm{M}] \\ 0 & \text{si } x \notin [x_\mathrm{m}, x_\mathrm{M}] \end{cases}$$

1. *Représentez graphiquement (sur le papier) $f_X(x)$. Vérifiez que $f_X(x)$ définit bien une densité de probabilité. Donnez l'expression analytique de la moyenne et de la variance.*

Dans un second temps, la variable X est distribuée selon une densité normale de moyenne m et de variance r. On note $\gamma = r^{-1}$ la précision. La densité de probabilité s'écrit :

$$f_X(x) = (2\pi)^{-1/2}\, \gamma^{1/2}\, \exp -\frac{1}{2}\gamma(x-m)^2 \,.$$

2. *Représentez graphiquement (sur le papier) $f_X(x)$. Commentez la manière dont la densité évolue avec m et r.*

1 Génération de réalisations d'une variable aléatoire

1.1 Densité uniforme

Le logiciel Matlab (ou Octave) dispose d'un générateur de nombres pseudo-aléatoires, la fonction `rand`, qui simule des « réalisations » de v.a. indépendantes uniformément distribuées sur $[0, 1]$ (voir l'aide de la fonction pour plus d'informations sur son utilisation).

3. *Tapez la commande* `rand` *dans la fenêtre de commande Matlab / Octave. Recommencez plusieurs fois de suite. A quoi correspondent les valeurs renvoyées ? Tapez également* `rand(1,3)`, `rand(3,2)`.

Nous allons utiliser cette fonction pour générer une suite de réalisations $x_1, x_2, \ldots x_N$ de la variable aléatoire X. Pour cela, créez un fichier de commandes et saisissez les lignes suivantes.

```
% Nettoyage
    close all, clearvars

% Parametres de simulation
    N=100;
    n=1:N;

% Realisations
    x=rand(1,N);

% Observations
    figure(1)
    plot(n,x,'*')
    title('Re''ealisations')
    grid
```

4. *Interprétez et commentez la figure obtenue.*

5. *Ajoutez une légende pour chacun des deux axes en utilisant les fonctions* `xlabel` *et* `ylabel`.

La probabilité que la v.a. X soit comprise entre $x - \delta/2$ et $x + \delta/2$ (avec δ petit) peut être approchée par l'aire du rectangle de hauteur $f_X(x)$ et de largeur δ, c'est-à-dire $f_X(x)\delta$:

$$\Pr\{\ X \in [x - \delta/2; x + \delta/2[\ \} \approx f_X(x)\delta\,.$$

Par ailleurs, à partir des N réalisations de la v.a., on peut approcher cette même probabilité par le rapport entre le nombre de points x_n situés dans l'intervalle $[x - \delta/2; x + \delta/2[$, noté N_x, et le nombre de points total N :

$$\Pr\{\ X \in [x - \delta/2; x + \delta/2[\ \} \approx \frac{N_x}{N}\,.$$

Par conséquent, une approximation de la densité de probabilité peut être obtenue par un histogramme normalisé de la suite de réalisations de la v.a. :

$$f_X(x) \approx \frac{N_x}{N\delta}\,.$$

6. *Tracez l'histogramme des réalisations à l'aide de la commande* `histogram` *(consulter la documentation :* `help histogram`*). Obtenez-vous une bonne approximation de la densité de probabilité ? Étudiez l'influence du nombre de réalisations et du nombre de classes de l'histogramme.*

7. *Calculez la moyenne et l'écart-type empiriques des N réalisations en utilisant les fonctions* `mean` *et* `std`. *Comparez avec la moyenne et la variance théoriques. Comment évoluent les choses avec N ? Concluez.*

8. *Transformation...*

8a. *Proposez une solution pour générer des réalisations uniformément distribuées sur les intervalles $[1, 2]$, puis $[0, 3]$ et enfin $[-1, 1]$.*

8b. *Vérifiez que vos propositions fournissent les résultats attendus.*

1.2 Densité normale

Le logiciel dispose également d'un générateur de nombres pseudo-aléatoires `randn` simulant des v.a. indépendantes distribuées sous une densité normale centrée et réduite (moyenne $m = 0$ et variance $r = 1$).

9. *Modifiez votre programme afin de générer une suite de réalisations selon cette densité.*

10. *Interprétez et commentez les deux figures (par* `plot` *et* `histogram`*) obtenues.*

11. *Comparez la moyenne et la variance empiriques des N réalisations avec les quantités théoriques. Que se passe-t-il lorsque vous augmentez ou réduisez N ?*

12. *Transformation...*

12a. *Proposez une solution afin de générer des réalisations d'une v.a. distribuée selon une densité normale avec : $m = 2$ et $r = 1$ puis $m = 0$ et $r = 3$ et enfin $m = 2$ et $r = 3$.*

12b. *Vérifiez que vos propositions fournissent les résultats attendus.*

1.3 Densités mélanges

Dans cette partie on s'intéresse à la notion de mélanges de deux densités. Ce sont des modèles qui apparaissent assez naturellement dans de nombreux problèmes pratiques. Fondamentalement, ils reposent sur deux variables.

1. Une variable d'étiquette L (pour Label en anglais) binaire décrivant un « statut ». On note arbitrairement $L = 1$ et $L = 2$ les deux valeurs possibles de la variable.
2. Une variable réelle notée X comme la variable uniforme ou la variable gaussienne précédemment.

Le modèle semble plus naturellement décrit sous forme hiérarchique : (1) une loi (marginale) pour L et (2) une loi (conditionnelle) pour X, dépendant de la valeur de L.

Concernant la variable L, elle suit une loi de Bernoulli de paramètre p :

$$L \sim \mathcal{B}(\ell\,;\,p) \quad \equiv \quad \Pr\{L = \ell\} = \begin{cases} p & \text{si } \ell = 1 \\ 1-p & \text{si } \ell = 2 \end{cases}$$

13. *Proposez une méthode pour simuler des échantillons de L à partir de la routine* `rand`*. Comme précédemment, tracez la suite des valeurs obtenues. Comment vérifier empiriquement que tout est en ordre ?*

Concernant la variable X, elle suit une distribution dépendant de la valeur de L :

$$\begin{cases} X \,|\, L = 1 & \sim & f_{X|L}(x \,|\, L = 1) = \dots \\ X \,|\, L = 2 & \sim & f_{X|L}(x \,|\, L = 2) = \dots \end{cases}$$

Par exemple, on peut imaginer
- deux gaussiennes, de paramètres (μ_1, γ_1) lorsque $L = 1$ et (μ_2, γ_2) lorsque $L = 2$,
- deux densités uniformes, l'une sur $[x^1_{\mathrm{m}}, x^1_{\mathrm{M}}]$ lorsque $L = 1$ et l'autre sur $[x^2_{\mathrm{m}}, x^2_{\mathrm{M}}]$ lorsque $L = 2$,
- ou même un mélange hétérogène...

On peut naturellement les simuler comme dans les questions précédentes.

14. *Travail du mélange.*

14a. *Proposez une méthode pour simuler des échantillons de X (au sens marginal du terme) à partir de la construction hiérarchique ci-dessus (simuler L puis $X \mid L$).*

14b. *Comme précédemment, tracez la suite des valeurs obtenues. Ce que vous observez vous semble-t-il correct ?*

14c. *Vérifiez que vous avez bien la « bonne » moyenne et la « bonne » variance.*

2 Couple, transformation linéaire et corrélation

Dans cette partie, on s'intéresse à la construction d'un couple de v.a. corrélées à partir d'un couple de variables décorrélées (et même indépendantes) par transformation linéaire.

15. *On considère un couple de v.a. (X_1, X_2) indépendantes distribuées selon une densité normale centrée et réduite.*

15a. *Générez $N = 1000$ réalisations de chacune des variables. Tracez dans le plan (X_1, X_2) les réalisations d'une v.a. en fonction des réalisations de l'autre (sans relier les points entre eux). Commentez le nuage de points obtenus.*

15b. *Quelle est la valeur du coefficient de corrélation du couple (X_1, X_2) ? Calculez numériquement une approximation de ce coefficient en utilisant la fonction Matlab `corrcoef` ou la fonction fournie `CalcCorCoef`. Comparez résultat théorique prévu et résultat numérique.*

15c. *Répétez plusieurs fois l'opération pour apprécier la variablité d'une réalisation à l'autre. Faites varier N et commentez la comparaison théorique - numérique.*

On considère maintenant deux v.a. obtenues à partir de (X_1, X_2) par la transformation linéaire suivante, avec $\varepsilon \in \mathbb{R}$:

$$\begin{aligned} Y_1 &= X_1 + \varepsilon X_2 \\ Y_2 &= X_2 + \varepsilon X_1 \end{aligned}$$

16. *Calculez les moyennes et les variances de Y_1 et de Y_2. Calculez ensuite le coefficient de corrélation $\rho(\varepsilon)$ du couple (Y_1, Y_2) en fonction de ε.*

17. *On fixe $\varepsilon = 2 + \sqrt{3}$.*

17a. *Quelle est la valeur de ρ correspondante ?*

17b. *Générez les $N = 1000$ réalisations de ce couple (Y_1, Y_2) à partir de celles du couple (X_1, X_2) générées à la question précédente. Tracez-les dans le plan (Y_1, Y_2) et comparez le nuage de points à celui obtenu à la question précédente.*

17c. *Calculez numériquement une approximation du coefficient de corrélation et comparez résultat théorique et résultat numérique. Ici encore : répétez l'opération à l'identique et faites varier N.*

18. *Que se passe-t-il pour $\varepsilon = 0$? Déterminez la ou les valeurs qui maximisent ou minimisent ce coefficient de corrélation. Donnez les valeurs de ε pour lesquelles ρ vaut $1/2$ ou $-1/2$. Déterminez les limites lorsque ε tend vers $\pm\infty$.*

19. *Question facultative : recommencez toute cette partie 2 avec une variable uniforme au lieu d'une variable gaussienne. Il suffit de remplacer `randn` par `rand`...*