L2 Mention Informatique

UE Probabilités

Chapitre 4 : Simulation - Régression

Notes de cours rédigées

par

Régine André-Obrecht, Julien Pinquier

I- Simulation de variables aléatoires

1. Introduction

Dans certaines expériences « réelles », où le hasard intervient, il peut être intéressant de réaliser des expériences numériques où on simule des variables aléatoires qui interviennent plutôt que de réaliser physiquement les expériences.

Exemple : Physique nucléaire → désintégrations des atomes fissibles décrits par de v. a. (loi exponentielle, loi de Poisson, etc.) avant de construire un réacteur !

John Von Neumann (Mathématicien, 1903-1957) : « Quiconque considère des méthodes arithmétiques pour produire des nombres aléatoires est, bien sûr, en train de commettre un péché ».

Il est, en général, exclu de réaliser matériellement des dispositifs fournissant des nombres aléatoires. Mais il existe des tables dites tables de nombres aléatoires et qui sont telles, que la suite des nombres qui y figurent est assimilable à la réalisation de tirages avec remise dans une urne de dix catégories de boules figurant à proportions égales.

Il existe aussi des algorithmes qui simulent (étant en fait déterministes) les nombres aléatoires.

2. Algorithmes générant des nombres pseudo-aléatoires

a) Méthode de Von Neumann

Méthode *middle-square* (carré médian). Très simple, elle consiste à prendre un nombre, à l'élever au carré et à prendre les chiffres au milieu comme sortie. Celle-ci est utilisée comme graine pour l'itération suivante.

b) Méthode de Fibonacci

Cette méthode est basée sur la suite de Fibonacci, modulo la valeur maximale désirée :
$x_n = (x_{n-1} + x_{n-2}) \bmod(M)$ avec x_0 et x_1 en entrée.

c) Méthode de Lehmer

On définit une suite d'entiers $x_0, x_1, \ldots, x_n$ ainsi :

- x_0 entier positif arbitraire,
- $x_{n+1} = k\, x_n \bmod(m)$, $n \geq 1$ avec m nombre premier.

Exemple : $m = 2^{31} - 1$ (nombre premier de Mersenne : 2^p-1 avec p nombre premier).

Il s'agit de la méthode utilisée par la fonction « rand » (loi uniforme) de Scilab et de Matlab (jusqu'à la version 4, mais encore disponible avec l'option 'seed').

d) Méthode de Mersenne Twister

Il s'agit de la méthode utilisée par la fonction « rand » (loi uniforme) de Matlab (depuis la version 7.4) et des langages C/C++.

3. Les tables de nombres aléatoires

Ces tables (cf. Figure 1) permettent d'extraire des entiers « aléatoires » dans l'intervalle de 0 à 10^k. On prend k colonnes, les chiffres dans chaque ligne forment un nombre de k chiffres. Après on prend les k colonnes suivantes, etc.

Pour passer aux nombres qui ne sont pas entiers, on peut diviser les nombres choisis par 10^k : on obtient les nombres dans l'intervalle [0, 1].

C'est une simulation de nombres aléatoires de loi uniforme.

TRENTE-CINQUIÈME MILLE

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	02 22	85 19	48 74	55 24	89 69	15 53	00 20	88 48	95 08	00 47
2	85 76	34 51	40 44	62 93	65 99	72 64	09 34	01 13	09 74	90 65
3	00 88	96 79	38 24	77 00	70 91	47 43	43 82	71 67	49 90	37 09
4	64 29	81 85	50 47	36 50	91 19	09 15	98 75	60 58	33 15	51 44
5	94 03	80 04	21 49	54 91	77 85	00 45	68 23	12 94	23 44	36 88
6	42 28	52 73	06 41	37 47	47 31	52 99	89 82	22 81	86 55	99 09
7	09 27	52 72	49 11	30 93	33 29	54 17	54 48	47 42	04 79	18 84
8	54 68	64 07	85 32	05 96	54 79	57 43	96 97	30 72	12 19	41 70
9	25 04	92 29	71 11	64 10	42 23	23 67	01 19	20 58	35 93	39 46
10	28 58	32 91	95 28	42 36	98 59	66 32	15 51	46 63	57 10	83 55
11	64 35	04 62	24 87	44 85	45 68	41 66	19 17	13 09	63 37	15 33
12	61 05	55 88	15 01	15 77	12 90	69 34	36 93	52 39	36 23	59 73
13	98 93	18 93	86 98	99 04	75 28	30 05	12 09	57 35	90 15	98 07
14	61 89	35 47	16 32	20 16	78 52	82 37	26 33	67 42	11 93	35 61
15	94 40	82 18	06 61	54 67	03 66	76 82	90 31	71 90	39 27	97 85
16	54 38	58 65	27 70	93 57	59 00	63 56	18 79	85 52	21 03	03 16
17	63 70	89 23	76 46	97 70	00 62	15 35	97 42	47 54	60 60	78 12
18	61 58	65 62	81 29	69 71	95 53	53 69	20 95	66 60	50 70	22 97
19	51 68	98 15	05 64	43 32	74 03	44 63	52 38	67 59	56 69	11 14
20	59 25	41 48	64 79	62 26	87 86	94 30	43 54	26 98	61 38	63 44
21	85 00	02 24	67 85	88 10	34 01	54 53	23 77	33 11	19 68	13 50
22	01 46	87 56	19 19	19 43	70 25	24 29	48 22	44 81	35 40	33 23
23	42 41	25 10	87 27	77 38	05 90	73 03	95 46	88 82	25 02	05 00
24	03 57	14 03	17 80	47 85	94 49	89 55	10 27	19 50	20 37	02 71
25	18 95	93 40	45 43	04 57	17 03	34 54	83 91	69 02	90 72	98 45

Figure 1: exemple de table de nombres aléatoires.

4. Simulation d'une variable aléatoire discrète

On cherche à définir une variable aléatoire discrète Y de loi :
$p_1 = P(Y = d_1)$, $p_2 = P(Y = d_2)$, …, $p_n = P(Y = d_n)$, …

Soit X une variable aléatoire de loi uniforme sur [0, 1].
On peut prendre une partition de [0, 1] en intervalles de longueurs p_1, p_2, …p_n, …
Par exemple, $[0, p_1[, [p_1, p_1 + p_2[, [p_1 + p_2, p_1 + p_2 + p_3[, \ldots$.

On peut définir une variable aléatoire Y' ainsi :

$Y' = d_1 \Leftrightarrow X \in [0, p_1[$,

$Y' = d_2 \Leftrightarrow X \in [p_1, p_1 + p_2[$,

$Y' = d_3 \Leftrightarrow X \in [p_1 + p_2, p_1 + p_2 + p_3[$,

...

La loi de Y' est égale à la loi de Y, donc des réalisations de Y' sont des réalisations de Y.

Exemple : Si la loi de Y est $p_1 = ½$, $p_2 = 3/8$ et $p_3 = 1/8$, alors la partition donne :

$[0,1/2\ [, [1/2,7/8[, [7/8,1[$.

5. Simulation d'une variable aléatoire réelle

On cherche à définir une variable aléatoire réelle Y de densité $f_Y(y)$.
Soit X une variable aléatoire réelle de densité $f_X(x)$.

On peut obtenir l'équation : $f_Y(y)\,dy = f_X(x)\,dx$.

Si X est de loi uniforme sur [0, 1], cette équation devient : $f_Y(y)\,dy = dx$.

Si on cherche y comme une fonction croissante de x, y = y(x), on peut écrire :

$F_Y(y(x)) = \int_{-\infty}^{y(x)} f_Y(y)\,dy = \int_0^x dx = x$, $0 \le x \le 1$ avec F_Y la fonction de répartition.

Si F^{-1} est la fonction inverse de F_Y alors $y(x) = F^{-1}(x)$.

Cela peut être utilisé pour calculer les nombres pseudo-aléatoires de loi non-uniforme à partir des nombres de la loi uniforme.

Exemple : Soit Y une variable aléatoire de loi exponentielle : $f_Y(y) = \lambda e^{-\lambda y}$.

$F_Y(y) = 0$ si $y < 0$,

$F_Y(y) = \int_0^y \lambda e^{-\lambda y} dy = \left[-e^{-\lambda y}\right]_0^y = 1 - e^{-\lambda y}$ sinon.

Il suffit alors de résoudre l'équation : $1 - e^{-\lambda y} = x$ et on obtient : $y = -\frac{1}{\lambda}\ln(1-x)$.

II- Régression linéaire - Approximation d'une variable aléatoire – Explication d'une variable aléatoire

1. Définition du problème

Considérons un vecteur de variables aléatoires $(Y, X_1, X_2, \ldots, X_n)$ supposées centrées :
$E(Y) = E(X_1) = E(X_2) = \ldots = E(X_n) = 0$.

Objectif : approximation de Y par une combinaison linéaire de $X_1, X_2, \ldots, X_n$.

Nous cherchons à écrire Y sous la forme : $Y = \sum_{i=1}^{n} a_i X_i + Y'$

où Y' est une erreur (déviation) non corrélée à $X_1, X_2, \ldots, X_n$.

Les coefficients du vecteur $A = (a_1, a_2, \ldots, a_n)$ sont appelés **coefficients de régression linéaire**.

Rappel : Covariance de 2 variables aléatoires U et V : $cov(U,V) = E(UV) - E(U).E(V)$.
Remarquons que $cov(U,U) = E(U^2) - E(U)^2 = var(U)$.

Supposons que var(Y), cov(Y, Xi) et $cov(X_i, X_j)$ existent pour tous $1 \leq i,j \leq n$.

Notons :

- $B = (E(Y, X1), E(Y, X2), \ldots, E(Y, Xn))$,
- B^t est le vecteur transposé de B,
- $\Gamma = \begin{pmatrix} \mathrm{cov}(X_1, X_1) & \mathrm{cov}(X_1, X_2) & \ldots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \ldots & \mathrm{cov}(X_n, X_n) \end{pmatrix}$.

La matrice $\begin{pmatrix} \mathrm{var}(Y) & B \\ B^t & \Gamma \end{pmatrix}$ est appelée **matrice de variance-covariance** de $(Y, X_1, X_2, \ldots, X_n)$.

Remarque : cette matrice est symétrique.

Pour chaque X_j, la condition de corrélation nulle (covariance nulle) de $Y' = Y - \sum_{i=1}^{n} a_i X_i$ et X_j a pour conséquence l'équation suivante :

$$cov(Y', X_j) = cov(Y, X_j) - \sum_{i=1}^{n} a_i cov(X_i, X_j) \Leftrightarrow B_j = \sum_{i=1}^{n} a_i cov(X_i, X_j)$$

car :

- $cov(Y', X_j) = 0$ (condition),
- $cov(Y, X_j) = E(YX_j)$ car $E(Y) = E(X_j) = 0$ (v. a. centrées).

Pour déterminer le vecteur $A = (a_1, a_2, \ldots, a_n)$, nous devons résoudre le système suivant :

$$\begin{cases} B_1 = \sum_{i=1}^{n} a_i cov(X_i, X_1) \\ B_2 = \sum_{i=1}^{n} a_i cov(X_i, X_2) \\ \quad \ldots \\ B_n = \sum_{i=1}^{n} a_i cov(X_i, X_n) \end{cases}.$$

Sous forme matricielle, ce système s'écrit ainsi : $B = A\Gamma$.

Si la matrice Γ est régulière (inversible), $A = B\Gamma^{-1}$. On a alors une solution et cette solution est unique.

Remarque : nous ne considèrerons que des cas de matrices régulières.

2. Cas de 2 variables centrées

Dans le cas de 2 variables centrées Y et X, la représentation que nous cherchons est :
$Y = a.X + Y'$.

La matrice de variance-covariance est de taille 2x2 : $\begin{pmatrix} var(Y) & cov(Y,X) \\ cov(X,Y) & var(X) \end{pmatrix}$.

Le système à résoudre s'écrit simplement : $cov(Y,X) = a.var(X)$.

Alors $a = \frac{cov(Y,X)}{var(X)}$ donc $Y = \frac{cov(Y,X)}{var(X)}X + Y'$.

Tenant compte du fait que la variance est égale à $var(X) = \sigma(X)^2$ et que le coefficient de corrélation est égal à $corr(X,Y) = \frac{cov(X,Y)}{\sigma(X).\sigma(Y)}$, on obtient : $a = \frac{\sigma(Y)}{\sigma(X)}corr(X,Y)$.

3. Cas de 2 variables non centrées

Dans le cas de 2 variables non centrées Y et X, la représentation de Y prend en compte les valeurs moyennes (espérances) de Y $m_Y = E(Y)$ et de X $m_X = E(X)$.

Alors : $Y = \frac{cov(Y,X)}{var(X)}(X - m_X) + m_Y + Y'$.

La droite d'équation : $\boxed{y = \frac{cov(Y,X)}{var(X)}(x - m_X) + m_Y}$ s'appelle la **ligne de régression** de Y sur X.

Cette droite a pour propriété que la variance var(Y') est minimale par rapport à toutes les décompositions de Y en partie linéaire (par rapport à X).

Unité de cours Probabilités et statistiques - Exercices
Chapitre 4 –Simulation et régression

Exercice 1 (Simulation)

Prendre les 4 premières colonnes de la table des nombres aléatoires (figure 1). Considérer les 4 chiffres comme les 4 décimales d'un nombre $X \in [0,1]$.

1- Faire un histogramme correspondant aux 20 premières lignes avec une partition de [0,1] en 4 intervalles de longueur 0,25.

2- En déduire les probabilités expérimentales correspondantes.

Exercice 2 (Simulation v. a. discrète)

Prendre les 10 premières lignes et les colonnes 5 à 8 de la table des nombres aléatoires.
Considérer les 4 chiffres de chaque ligne comme les 4 décimales d'un nombre $X \in [0,1]$.

1- Simuler une variable aléatoire $Y \in \{0, 1, 2\}$ de loi :

$$P(Y = 0) = \frac{1}{3},\ P(Y = 1) = \frac{1}{6} \text{ et } P(Y = 2) = \frac{1}{2}.$$

2- Comparer les probabilités pratiques obtenues aux probabilités théoriques attendues.

Exercice 3 (Simulation v. a. réelle)

Soit une variable aléatoire réelle X de loi uniforme sur [0,1]. Soit une v. a. réelle Y de densité :

a) $f(y) = 2y$ pour $0 \le y \le 1$
0 sinon,

b) $f(y) = y$ pour $0 \le y \le 1$
$2-y$ pour $1 \le y \le 2$
0 sinon,

c) $f(y) = (k+1)\, y^k$ pour $0 \le y \le 1$,
0 sinon.

Trouver, pour chacun des 3 cas, une fonction $y = h(x)$ telle que $Y = h(X)$.

Exercice 4 (Simulation et régression)

Prendre 2 réalisations (X et Y) d'échantillons de loi uniforme de taille 10 (les colonnes 3 et 5 des 10 premières lignes de la table des nombres au hasard).

1- Estimer la loi empirique conjointe et les lois marginales de X et Y.

2- Faire les histogrammes des lois marginales X et Y en prenant comme intervalles :
[-0,5 ; 0,5], [0,5 ; 1,5], etc.

3- Trouver l'équation de la ligne de régression de Y par rapport à X.

Exercice 5 (Régression)

Soient X et Y, deux variables aléatoires indépendantes de loi uniforme qui prennent comme valeurs 1, 2 et 3. On pose $Z = X + Y$ et $U = XY$.

Reprendre les résultats de l'exercice 3 (cf. chapitre 2) afin de trouver l'équation de régression linéaire de U sur Z.

Exercice 6 (Régression)

Soit la loi conjointe du couple (X, Y) donnée par :

X \ Y	-2	-1	1	2
-2	1/8	0	1/8	0
-1	0	1/8	1/8	0
1	0	0	0	1/4
2	0	0	1/4	0

Calculer l'équation de la ligne de régression de Y par rapport à X.

Exercice 7 (Régression et corrélation)

Soit (X, Y) un couple de variables aléatoires de loi uniforme dans le triangle ABC avec les coordonnées suivantes : A (0,0) – B (1,1) – C (1,0).

1- Estimer la densité jointe et les densités marginales.

2- Calculer les espérances, les variances et le coefficient de corrélation.

3- Trouver l'équation de régression de Y sur X.