Séances n°4 et 5 : simulation d'une loi non uniforme et test de conformité

Module de Génie Informatique

Laboratoire LTSI - UMR INSERM 642 - Université de Rennes1

I - Génération de Variables aléatoires par la méthode d'inversion

I.1 -Le problème

Sur calculateur, dans beaucoup de langages de programmation, on dispose d'une primitive notée **rand** (rand est le raccourci de random qui signifie aléatoire en anglais) telle que l'exécution du segment du code (écrit dans un style matlab,% correspondant à une ligne de commentaire):

```
% T : tableau de N réels indicé de 1 à N
for i=1:N
X(i)=rand
end
```

aboutit à ce que les N nombres stockés dans les N variables X(i) (une variable correspond à une case mémoire en machine) sont assimilables aux résultats de N tirages aléatoires **indépendants** suivant une même **loi uniforme sur [0,1].**

Or dans l'application des méthodes de simulation à l'analyse de problèmes réels, on doit considérer que les VA introduites dans la phase de modélisation (on doit toujours modéliser avant de simuler) suivent des lois éventuellement très différentes de la loi uniforme simulable par la primitive rand.

Vous avez ainsi vu dans le TP1 comment simuler une loi discrète avec rand et comment simuler une loi uniforme sur un intervalle quelconque [a,b] en exécutant X= (a+b)/2+ (b-a)*(rand-1/2) où :

- rand-1/2→loi uniforme sur [-1/2,1/2]
- (b-a)*(rand-1/2)→loi uniforme sur [-(b-a)/2,(b-a)/2]
- X= (a+b)/2+ (b-a)*(rand-1/2→ loi uniforme sur [-(b-a)/2+(a+b)/2,(b-a)/2+(a+b)/2]=[a,b]

Ici nous nous posons la question de simuler une **loi de type continu** pour laquelle on se donne soit la fonction de répartition $F(x), x \in R$, soit la densité de probabilité $p(x), x \in R$. Rappelons que $p(x) = \frac{d}{dx} F(x), \int_{-\infty}^{x} p(u)du = F(x), x \in R$ et que F est toujours une fonction non décroissante, tendant vers 0 quand x tend vers $-\infty$ et vers 1 quand x tend[i] vers $+\infty$ et qui est de plus continue.

I.2 -Une solution : la méthode d'inversion .

1) Rappelons d'abord comment obtenir par le calcul la loi de probabilité d'une VA Y de la forme $X = f(U)$, U étant une VA de loi continue.

Supposons ici que *f* est strictement croissante (en plus d'être continue). Cette fonction admet donc une inverse notée f^{-1} telle que $f^{-1}(f(x)) = f(f^{-1}(x)) = x, x \in R$ et qui est elle même continue et strictement croissante. Il est alors facile de trouver la loi de *X*. On a :

$F_X(x) = P(X < x) = P(f(U) < x) = P(f^{-1}(f(U)) < f^{-1}(x)) = P(U < f^{-1}(x))$ et donc

$\boxed{F_X(x) = F_U(f^{-1}(x))}$ (1)

On trouvera en annexe les relations permettant d'obtenir la densité de probabilité correspondante.

2) Considérons à présent que la VA *U* est de loi uniforme sur l'intervalle]0,1[et que $F : R \to]0,1[$ est la fonction de répartition ciblée, supposée continue et strictement croissante ce qui entraîne que l'inverse $F^{-1} :]0,1[\to R$ existe. Considérant la loi de probabilité de la VA *X* définie par $X = F^{-1}(U)$ et d'après le résultat (1) ci dessus on a

$F_X(x) = F_U((F^{-1})^{-1}(x)) = F_U(F(x)) = P(U < F(x))$

et, puisque $F(x)$ est une valeur dans]0,1[et que *U* est de loi uniforme sur (0,1) on aboutit à :

$\boxed{F_X(x) = P(U < F(x)) = F(x), x \in R}$

Donc l'instruction $X = F^{-1}(RAND)$ fait prendre à *X* une valeur aléatoire qui suit une loi de probabilité admettant pour fonction de répartition la fonction *F* ciblée.

3) Application à la loi exponentielle.

La loi exponentielle est fréquemment rencontrée en pratique pour simuler des instants d'arrivée ou des durées aléatoires. On rappelle qu'une VA de loi exponentielle est à valeurs réelles positives, de loi continue avec une densité de la forme $p(x) = ae^{-ax}1_{[0,\infty[}(x), x \in R$ où *a* est un paramètre réel positif. La fonction de répartition correspondante est $F(x) = (1 - e^{-ax})1_{[0,\infty[}, x \in R$. En observant que la restriction sur les réels positifs de cette fonction, $F : R^{*+} \to]0,1[$ est continue strictement croissante et admet une inverse $F^{-1} : [0,1[\to R^{*+}$ on peut trouver son inverse en résolvant en *u* sur R^{*+} l'équation $u = F(x) = 1 - e^{-ax}$ dont la solution est $x = \frac{-1}{a}\log(1-u)$ et où *x* prend ses valeurs dans $]0,\infty[$ quand *u* prend les siennes dans]0,1[. On en déduit que l'instruction $X = -\frac{1}{a}\log(1 - RAND)$ fait prendre à *X* une valeur aléatoire suivant la loi exponentielle de paramètre *a*.

4) Exercice1 : Vérifier que si *U* est une VA de loi uniforme sur]0,1[alors $U_1 = 1 - U$ est encore une VA de loi uniforme sur]0,1[. En déduire une légère simplification de l'instruction ci dessus pour obtenir *X* de loi exponentielle.

I.3 - La vérification expérimentale

Nous allons considérer ici la fonction de répartition dite expérimentale. Cette dernière est notée ici *FE* et est définie comme suit :

$$FE(x) = \frac{1}{N_e}\sum_{1}^{N_e} H(x - x_i), x \in R$$

où *H* est la fonction échelon unité **continue à gauche (voir remarque ci dessous)** et où $x_1,..,x_{N_e}$ sont N_e réalisations indépendantes obtenues par autant d'exécutions de $X = F(rand)$. Ainsi, $FE(x)$ a la forme d'une fonction en escalier avec des marches de hauteur 1/Ne en supposant, *ce que l'on fera dans la suite*, que les xi sont toutes distinctes).

Remarque : dans la définition $F_X(x) = P(X < x)$ l'inégalité stricte implique la continuité à gauche alors que si la définition prise était $F_X(x) = P(X \le x)$ on aurait une continuité à droite.

Pour étudier pratiquement *FE* à partir des données on ordonne l'ensemble $\{x_1,..,x_{N_e}\}$ (au moyen de la primitive sort en matlab) pour obtenir l'ensemble $\{y_1,..,y_{N_e}\}$ contenant les mêmes valeurs mais où $k > l \Rightarrow y_k \ge y_l$. Il est bien entendu important de pouvoir visualiser (en utilisant plot() en matlab) les fonctions de répartitions théorique et expérimentale. Pour cela, on devra discrétiser l'ensemble des valeurs de x en ne retenant qu'un nombre fini d'abscisses jugées intéressantes. Typiquement on choisira judicieusement des valeurs extrêmes $x_{\min}$ et $x_{\max}$ et un nombre $n_a + 1$ d'abscisses sur $[x_{\min}, x_{\max}]$ de valeurs $\alpha_k = x_{\min} + k\frac{(x_{\max} - x_{\min})}{n_a}, k = 0,..,n_a$. Les valeurs théoriques de la fonction de répartition seront bien entendu les valeurs $F(\alpha_k)$ tandis que les valeurs de la fonction de répartition expérimentale seront égales à :

$FE(\alpha_k) = \frac{l}{Ne}$ $y_l < \alpha_k \le y_{l+1}, l = 1,..,N_e - 1$

$FE(\alpha_k) = 0$ si $\alpha_k \le y_1$, $FE(\alpha_k) = 1$ si $\alpha_k > y_{N_e}$

I.5 Travail demandé

1) Générer et stocker 1000 réalisations avec le code $X = -\frac{1}{0.2}\log(rand)$ qui, d'après la théorie, correspond à une loi exponentielle de paramètre 0.2.

2) Calculer la moyenne expérimentale et la variance expérimentales sur les données obtenues en 1) et comparer avec la moyenne et la variance théoriques attendues pour une loi exponentielle de paramètre 0.2, ce qui est une première indication quand à la conformité de ces données à une loi exponentielle de paramètre 0.2.

3) Générer et stocker 1000 réalisations avec le code $X = 5 + (rand - 1/2) * 5 * 12\text{^}0.5$ et reprendre 2) Que pensez vous du résultat ?

4) Ecrire un programme matlab pour représenter graphiquement une fonction de répartition exponentielle de paramètre donné et la fonction de répartition expérimentale pour une suite de valeurs observées. Utiliser ce code pour comparer les fonctions de répartition expérimentales des données générées en 1) et 2) avec la fonction de répartition théorique correspondant a) à la loi exponentielle introduite et b) à la loi des VA générées en 3). Expérimentez avec différentes tailles d'échantillon, par exemple N_e =20,50,100,500,1000

5)On considère une densité de probabilité $p(x) = x/4$ si $0 \le x \le 2$, $= 1 - x/4$ si $2 \le x \le 4 = 0$ sinon. Calculer la moyenne et la variance et trouver la fonction de répartition *F* correspondante.

Générer des réalisations avec le code $X = F^{-1}(rand)$, calculer la moyenne et la variance expérimentale pour comparer avec les valeurs théoriques, et enfin comparer visuellement les fonctions de répartition expérimentale et théorique.

II - Test de conformité d'une loi de probabilité expérimentale à une loi théorique.

II.1 - Le problème

Supposons que l'on dispose d'une fonction sur calculateur, développée dans le but de procéder à certaines simulations et qui est de la forme :

X=f(rand)
ou encore de la forme :
U1=rand ;…. ;Un=rand ; X=f(U1,..,Un),

et que l'on s'attende, d'après des raisonnements où des calculs théoriques, à ce que la loi de X corresponde à une fonction de répartition notée *F* (par exemple une loi exponentielle, une loi de Gauss,..), et que l'on appellera ici fonction de répartition théorique. Se pose alors la question de vérifier expérimentalement cette hypothèse, que l'on appellera 'hypothèse nulle' et que l'on notera H_0, à partir d'une suite d'échantillons de X, tirés indépendamment en activant N_e fois la fonction qui retourne X pour obtenir N_e valeurs au moyen du code suivant :

```
% X : tableau X(1)…X(Ne) de réels
% U : tableau U(1),..,U(n) de réels
for i=1:Ne
  for k=1:n
   U(k)=rand;

  end
  X(i)=f(U(1),..,U(n));
end
```

Une première manière répondre à la question ci-dessus est de considérer, comme nous l'avons fait dans la partie I, les représentations graphiques des fonctions de répartition expérimentale et théorique et de les comparer visuellement. Ceci peut cependant être considéré comme une méthodologie insatisfaisante sur le plan qualitatif, et peu convaincante quand Ne n'est pas très élevé.

II.2 – Test de conformité

Une approche plus 'carrée' consiste à introduire une procédure qui admet en entrée les valeurs X(1),..,X(Ne) et qui en sortie fournit une décision : soit l'acceptation, soit le refus de H0. Une telle procédure sera appelée **test de conformité** des observations à la loi théorique correspondant à H0.

Pour cela Kolmogorof a proposé de comparer quantitativement la fonction de répartition théorique attendue F à la fonction de répartition expérimentale au moyen d'une distance définie comme suit :

$d(F,FE) = \sup_{x\in R}\left|F(x) - FE(x)\right|$ (2),

et qui mesure l'écart maximal, quand *x* parcourt *R*, entre les valeurs de la fonction de répartition théorique et celles de la fonction de répartition expérimentale (la fonction en escalier).

Fixant une valeur d0>o le test de conformité de Kolmogorof est le suivant:

$d(F,FE) > d_0 \rightarrow H_0$ refusée,

$d(F,FE) \le d_0 \rightarrow H_0$ acceptée,

où d_0 étant un seuil positif à choisir par l'expérimentateur (voir plus loin).

Pour obtenir pratiquement $d(F,FE)$ à partir des données on peut montrer qu'il suffit de calculer :

$\max_{k=1,..,N_e}\left(\left|FE(y_{k+1}) - F(y_k)\right|, \left|F(y_k) - FE(y_k)\right|\right)$

Exercice 2 : montrer que la formule ci-dessus donne bien le plus grand écart possible tel qu'il est défini plus haut par la formule (2), ce qui indique qu'il suffit de considérer les valeurs de FE et F aux seules abscisses $y_k, k = 1,..,N_e$.

Il peut être démontré (cela est assez facile) que, pour Ne fixé, et quand l'hypothèse H0 est vraie, la variable aléatoire $D = d(F, FE(X_1,..,X_{N_e}))$ suit une loi de probabilité qui est toujours la même quelle que soit la fonction de répartition théorique. Dans ces conditions, en supposant ainsi que l'on est capable de calculer $P(D \geq s)$ pour N_e et s fixés et pour n'importe quelle loi théorique, le test suivant a été introduit par Kolmogorov :

- On 'collecte' les données $\{x_1,..,x_{N_e}\}$ et on calcule la distance $d(F, FE(x_1,..,x_k))$
- On calcule $P(D \geq d_0)$ avec $d_0 = d(F, FE(x_1,..,x_{N_e}))$
- Si $P(D \geq d_o)$ est *suffisamment petite* (par exemple <0.01) on rejette H0 sinon on l'accepte.

On a ainsi une procédure qui n'a rien d'idéal mais qui satisfait l'intuition. En effet, plus la valeur de $P(D \geq d_0)$ est petite moins il est vraisemblable que *D* ait pu prendre par hasard la valeur d_0 sous l'hypothèse H_0, ce qui amène à la rejeter. La question centrale en pratique est évidemment le choix de d_0 où, de manière équivalente, de $p_0 = P(D \geq d_0)$. Cependant, sans faire plus d'hypothèses sur le type de 'déformation' ,par rapport à la loi prévue dans H_0, de la loi effective dont sont issues les valeurs observées, la réponse à cette question ne peut pas s'appuyer sur de simples considérations mathématiques. C'est en quelque sorte à l'expérimentateur de 'prendre ses responsabilités', en tenant compte du contexte, quand au choix de 'la barre' sur p_0 en deçà de laquelle H_0 sera refusée.

La valeur de $P(D \geq d)$ en fonction de d peut être approchée au moyen de formules d'approximation. Ainsi Kolmogorov a montré que pour N_e grand :

$$P(D\sqrt{N_e} \geq \lambda) \simeq 1 - \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2\lambda^2)$$

Cependant il est également possible d'obtenir une bonne approximation des valeurs de $P(D \geq d_0)$ au moyen d'une simulation mettant en œuvre des tirages indépendants dans une loi quelconque (par exemple uniforme) , puisque, comme signalé plus haut, la loi de *D* ne dépend pas de la loi théorique prise en compte dans H_0 :

```
%Le code ci-dessous permet d'approximer  P(D>d0) pour Ne fixé
%L'approximation utilise Ns simulations faisant appel chacune à Ne appels à Rand
p=0 ;
for k=1:Ns

    X=rand(1,Ne) ;
    Y=sort(X) ;
    ---------------------------------------------------------------------------------------------
    %valeurs à gauche et à droite  de la fonction de répartition expérimentale
    FED= [1:Ne]/Ne ;
    FEG=[0 :Ne-1] /Ne ;
    ---------------------------------------------------------------------------------------------
    %fonction répartition théorique F(x)=x sur (0,1)
    F=Y ;
    ---------------------------------------------------------------------------------------------
    D=max(max( |FED-F |FEG-F | ) ;
    ---------------------------------------------------------------------------------------------
    %comptage des dépassements
    p=p+(D>=d0)
    ---------------------------------------------------------------------------------------------
end ;
pestim=p/Ns
---------------------------------------------------------------------------------------------
%Les appels rand(1,Ne)  peuvent être remplacés par Ne tirages indépendants dans une
%loi quelconque sans changer la valeur théorique de p0
---------------------------------------------------------------------------------------------
```

Il sera bien entendu toujours utile de visualiser sur un même graphe *F* et *FE* pour apprécier 'à l'œil' la proximité entre les 2 fonctions de répartition.

II.3 Travail demandé

1) Ecrire un programme matlab pour calculer la distance $d(F, FE)$ entre une fonction de répartition théorique et une fonction de répartition expérimentale.

2) Utiliser 1) dans le cas I.5 1) pour différentes valeurs de N_e en relançant , à N_e fixé, la simulation et le calcul de la distance pour apprécier qualitativement comment la VA $D = d(F, FE)$ se comporte.

3) Ecrire un programme matlab pour calculer de manière approchée par simulation la probabilité de dépassement $P(d(F, FE) \geq d_0)$ à N_e fixé. Utiliser ce programme pour vérifier effectivement que cette probabilité est plus faible lorsque *FE* correspond à des données qui n'ont pas été tirées dans la loi théorique de référence, en utilisant les situations de simulation introduites en I.5 1) et 3). Ceci devrait permettre d'affirmer que les données tirées dans la loi uniforme ne sont pas conformes à la loi théorique exponentielle.

4) Considérons la situation où H_0 correspond à une loi exponentielle mais de paramètre a non connu. Dans ce cas il est possible, sachant que la moyenne pour une telle loi est 1/*a*, de prendre comme loi de référence une loi exponentielle dont le paramètre est pris égal à l'inverse de la moyenne expérimentale des valeurs de l'échantillon. Ecrire un programme pour tester ce procédé et apprécier si dans ce cas la valeur $P(d(F, FE) \geq d_0)$ est notablement modifiée à N_e et p_0 fixés.

Annexe :

Calcul de la densité de probabilité de $X = f(U)$

$$p_X(x) = \frac{d}{dx} F_X(x) = F_U^{'}(f^{-1}(x)) \frac{d}{dx}(f^{-1}(x)) = p_U(f^{-1}(x)) \frac{1}{f^{'}(f^{-1}(x))}, x \in R$$

Si *f* est continue strictement décroissante on a

$$F_X(x) = P(X < x) = P(f(U) < x) = P(f^{-1}(f(U)) > f^{-1}(x)) = P(U > f^{-1}(x)) = 1 - F_U(f^{-1}(x))$$

(en tenant compte du fait, pour la dernière égalité, que $P(U = f^{-1}(x)) = 0$ puisque, la loi de *U* étant continue, *la probabilité pour que U tombe exactement sur une valeur particulière comme* $f^{-1}(x)$ *est nulle quelle que soit cette valeur si bien que* $1 - F_U(u) = P(U \geq u) = P(U > u)$). Par dérivation on obtient alors $p_X(x) = \frac{d}{dx} F_X(x) = -F_U^{'}(f^{-1}(x)) \frac{d}{dx}(f^{-1}(x)) = -p_U(f^{-1}(x)) \frac{1}{f^{'}(f^{-1}(x))}, x \in R$.

Remarquant que *f* est partout décroissante et que $\forall x : f^{'}(x) < 0$ on aboutit à une formule vraie chaque fois que *f* est <u>continue strictement monotone</u>, qu'elle soit <u>croissante ou décroissante</u> :

$$\boxed{p_X(x) = p_U(f^{-1}(x)) \frac{1}{\left| f^{'}(f^{-1}(x)) \right|}, x \in R}$$

Cette formule se généralise dans le cas où *f* est toujours continue dérivable mais où on ne suppose plus le caractère monotone. Elle devient :

$$p_X(x) = \sum_{u_k} p_U(f_k^{-1}(x)) \frac{1}{\left| f_k^{'}(f_k^{-1}(x)) \right|}, x \in R$$

u_k : solutions de l'équation $f(u) = x$

à condition de supposer toutefois que *f'* <u>ne s'annule qu'en des points isolés</u>.

---

[i] La limite pouvant être atteinte