TPs 2 et 3 : Chaines de Markov

Introduction : Les chaînes de Markov donnent un exemple de suites $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires non indépendantes, ce qui est le cas dans beaucoup de situations.
Elles ont des propriétés très intéressantes, notamment la convergence, dans certains cas, vers un "état d'équilibre" (ou "état stable"), et ceci quelles que soient les conditions initiales.

1. Présentation théorique des chaînes de Markov

a) Exemple introductif : épidémiologie

On modélise l'évolution d'une maladie au cours du temps en classant les individus en trois groupes : S (individus sains et non immunisés), I (individu immunisés) et M (individus infectés). $E = \{S, I, M\}$ est appelé l'ensemble des **états** possibles.
On donne la loi d'évolution suivante :

- à l'instant 0 (avant le début de l'épidémie), tous les individus sont dans l'état S.
- on constate que la moitié des individus appartenant au groupe S à l'instant n (n-ième jour) restent dans ce groupe à l'instant $n+1$ ((n+1)-ième jour), et que l'autre moitié tombe malade ; 10% des individus immunisés à l'instant n passent dans le groupe S à l'instant $n+1$, les autres restent immunisés ;1 malade sur 5 le reste, les autres guérissent et deviennent immunisés.

On choisit un individu au hasard dans la population, on note, pour tout $n \in \mathbb{N}$, S_n l'évènement "l'individu est sain et non immunisé à l'instant n" et $s_n = P(S_n)$, I_n l'évènement "l'individu est immunisé à l'instant n" et $i_n = P(I_n)$, et M_n l'évènement "l'individu est malade à l'instant n" et $m_n = P(M_n)$.

1. Représenter le modèle épidémiologique par un graphe (*appelé diagramme de transition*).
2. Exprimer en notations mathématiques les probabilités conditionnelles données par l'énoncé .
3. Enoncer la formule des probabilités totales sur le SCE $\{S_n, I_n, M_n\}$.
4. Donner les relations entre s_{n+1}, i_{n+1}, m_{n+1} et s_n, i_n, m_n.
5. On définit le vecteur colonne $U_n = \begin{pmatrix} s_n \\ i_n \\ m_n \end{pmatrix}$.
   Expliciter la matrice A telle que $U_{n+1} = AU_n$. Préciser U_0.
   *U_0 est appelé la **loi initiale** et A est appelée la **matrice de transition** de la chaîne de Markov.*
6. Justifier que : $\forall n \in \mathbb{N}$, $U_n = A^n U_0$.

b) Caractérisation par la matrice de transition et la loi initiale

i) Définition

Soit E un ensemble d'états (dans l'exemple, $E = \{S, I, M\}$).
On s'intéresse à une suite (X_n) de variables aléatoires telle que X_n désigne l'état d'un système à l'instant n, le système étant soumis à une loi d'évolution aléatoire à l'intérieur de l'ensemble d'état E.
Dans l'exemple épidémiologique, X_n représente l'état d'un individu au cours du temps et peut prendre les trois valeurs : S, I et M.

Définition 1

*Lorsque la loi de X_{n+1} (le futur) ne dépend que de l'état X_n (présent) (indépendamment de ceux qui l'on précédé i.e indépendamment du passé), on dit que X_n est une **chaîne de Markov**. Un processus de Markov est aussi dit "processus sans mémoire".*

Remarque

Les chaînes de Markov sont le pendant aléatoire des suites récurrentes d'ordre 1 $x_{n+1} = f(x_n)$.
Le futur (instant $n+1$) ne dépend que du présent (instant n), indépendamment du passé (instants $(x_0, x_1, ..., x_{n-1})$).
La différence avec les suites récurrentes étant que X_{n+1} est une fonction aléatoire de X_n.

Les chaînes de Markov sont un objet essentiel des probabilités modernes ; elles sont utilisées avec succès en théorie des jeux, physique (ex : modèle des Urnes d'Ehrenfest, 1996), biologie et épidémiologie, sciences sociales, finance, informatique.

ii) Matrice de transition

Définition 2

La matrice A de l'exemple est appelée ***matrice de transition****.*
C'est la matrice des probabilités conditionnelles :

$$A = \begin{pmatrix} P(S_{n+1}|S_n) & P(S_{n+1}|I_n) & P(S_{n+1}|M_n) \\ P(I_{n+1}|S_n) & P(I_{n+1}|I_n) & P(I_{n+1}|M_n) \\ P(M_{n+1}|S_n) & P(M_{n+1}|I_n) & P(M_{n+1}|M_n) \end{pmatrix}$$

On ne s'intéressera qu'à des chaînes de Markov dites ***homogènes****; c'est-à-dire pour lesquelles ces probabilités conditionnelles (et donc la matrice de transition) ne dépendent pas de n. Intuitivement, cela signifie que la "loi d'évolution" d'un instant à l'instant suivant reste toujours la même au cours du temps.*
$P(S_{n+1}|I_n)$ est alors appelée ***probabilité de transition de l'état I à l'état S****. Et on peut dessiner le* ***diagramme de transition*** *de la chaîne, dont les sommets sont les états et les flèches représentent les probabilités de transition entres ces états.*

Propriété 1

Les coefficients d'une matrice de transition sont positifs et la somme de chaque colonne d'une matrice de transition vaut 1.
On dit que c'est une **matrice stochastique**.

iii) Loi de l'état à l'instant n

Attention ! Les propriétés ci-dessous devront être redémontrées dans un sujet de concours

Propriété 2

Si on note U_n la matrice colonne de la loi de X_n, ici, $U_n = \begin{pmatrix} s_n \\ i_n \\ m_n \end{pmatrix}$, alors la formule des probabilités totales sur le SCE constitué de l'ensemble des états donne :

$$U_{n+1} = AU_n$$

Corollaire 1

On en déduit, par récurrence sur n, que pour tout $n \in \mathbb{N}$,

$$U_n = A^n U_0$$

Cette formule nous permet de calculer la loi de X_n à chaque instant n.
On remarque donc que les vecteurs U_n (les lois des X_n) sont entièrement déterminés par la donnée de la matrice A (*la matrice de transition*) et du vecteur colonne U_0 (*la loi initiale*).

Remarque

La matrice colonne $U_0 = \begin{pmatrix} s_0 \\ i_0 \\ m_0 \end{pmatrix}$ a ses coefficients positifs et de somme égale à 1. On dit que c'est un **vecteur stochastique**.

Exercice 1. : simulation de l'exemple épidémiologique
En informatique, on représente l'état S par le nombre 1, l'état I par le nombre 2 et l'état M par le nombre 3.

1. Compléter le programme (compléter les ***) suivant (ne pas le taper !!) pour qu'il simule la suite des états $X_0, X_1, ...X_n$ de l'exemple épidémiologique et les mémorise dans un vecteur X .
   Remarque : en informatique, un "vecteur" est une matrice ligne.

```
n=input('Entrer la durée de l'expérience :' )
X=zeros(1,n+1)
X(1)=1
for k=1:n do
    if X(k)==1 then
        if rand() <1/2 then
            X(k+1)=***
        else
            X(k+1)=***
        end
    elseif X(k)==2 then
        if *** then
            X(k+1)=***
       else
           X(k+1)=***
        end
    else
        if *** then
            X(k+1)=***
       else
           X(k+1)=***
        end
    end
end
disp(X)
```

2. La simulation des chaînes de Markov étant rapidement très lourde (cf ci-dessus pour seulement trois états), une fonction a été prédéfinie en Scilab pour la simuler :
- Soit $x0$ un état fixé ;
- Soit A la matrice des probas de transition de la chaîne de Markov ; L'instruction `X=grand(n,'markov',A',x0)` donne un vecteur ligne X représentant la simulation des variables $(X_1, X_2, ..., X_n)$ d'une chaîne de Markov de matrice de transition A et de position initiale $x0$. *Attention :*
-*On doit donner A' la transposée de A car la définition d'une matrice de transition en Scilab est la transposée de notre définition.*
- *Si on souhaite affiche l'état initial de la chaîne de Markov, il faut rajouter la valeur initiale $x0$ au vecteur X !*

   Simuler plusieurs fois la chaîne de Markov X de l'exemple épidémiologique pour $n = 10$ et vérifier la cohérence des résultats.

3. **Loi empirique de X_{10} :**

   (a) Compléter le programme suivant qui simule 100 chaînes de Markov $(X_1, X_2, ..., X_{10})$ du modèle épidémiologique et enregistre dans un vecteur p la liste des états de santé X_{10} constatés au 10 ième jour :

   ```
   n=100
   p=zeros(1,n)
   for k=1:n do
       X=grand(*****,'markov',A',*******)
       p(k)=************
   end
   disp(p, "p=")
   ```

   Exécuter le programme. On a ainsi obtenu un échantillon de 100 simulations de X_{10}.

   (b) Taper l'instruction `M=tabul(p,'i')`. Que fait l'instruction `tabul` ? Afficher la liste des modalités de X_{10} dans un vecteur `x` et la liste des fréquences associées dans un vecteur `f`.
   Recommencer avec 1000 simulation de X_{10} et noter le tableau des fréquences obtenues.

4. **Loi théorique de X_{10} :** On sait d'après le corollaire 1 que le vecteur U_k de la loi théorique de X_k est défini par la donnée de U_0 et la relation de récurrence $U_{k+1} = AU_k$. Compléter le programme suivant qui crée la matrice, de taille 3*11, $E = [U0 \quad U1 \quad ... \quad Un]$.

   ```
   // On crée une matrice [U0 U1 ...Un]
   E= zeros(3,11)
   ```

```
E(:,1)=[1;0;0]                    // range dans la première colonne de E la loi U0.
for i=1:10 do
   E(:,i+1)=*************     // range dans la i+1-ième colonne de E la loi U(i)
end
disp(E)
```

Lancer le programme et donner le tableau de la loi théorique de X_{10}.
a) Ce tableau est-il cohérent avec les résultats précédemment obtenus ?
b) En regardant la progression des colonnes, que voit-on apparaître ?

5. **Représentation graphique des lois :** Taper :

```
scf(1)
n=0:10
plot2d(n', [E(1,:)', E(2,:)',E(3,:)'])
title('les suites (sn), (in) et (mn)')
legend(['(sn)';'(in)'; '(mn)'])
```

On remarque alors que la loi de (X_n) converge vers une loi limite π.
En pratique, cela signifie que le pourcentage d'individus sains (respectivement immunisés, malades) se stabilise autour d'une valeur limite au bout d'un grand nombre de jours.

c) Convergence vers l'état stable

i) Etat stable

Propriété 3

Supposons qu'à un moment donné n_0, le vecteur $U_{n_0} = \begin{pmatrix} s_{n_0} \\ i_{n_0} \\ m_{n_0} \end{pmatrix}$ vérifie :

$$AU_{n_0} = U_{n_0}$$

i.e. Le vecteur U_{n_0} est un vecteur propre de A associé à la valeur propre 1 et un vecteur stochastique.

Alors, par récurrence sur $n \geq n_0$, on montre :

$$\forall n \geq n_0,\ U_n = U_{n_0}$$

Les suites (s_n), (i_n) et (m_n) sont donc constantes : la proportion d'individus dans chacun des états ne bouge pas. La distribution de probabilité (s, i, m) est alors appelée **probabilité stationnaire** ou **état stable**.
Un état stable, s'il existe, est unique (admis).

ii) Convergence vers l'état stable

Propriété 4

Plaçons-nous dans le cas où les suites (s_n), (i_n) et (m_n) convergent.

Notons $s = \lim\limits_{n\to+\infty} s_n$, $m = \lim\limits_{n\to+\infty} m_n$ et $i = \lim\limits_{n\to+\infty} i_n$. et Π le vecteur $\Pi = \begin{pmatrix} s \\ i \\ m \end{pmatrix}$.

Alors la relation $U_{n+1} = AU_n$ donne, par passage à la limite :

$$\Pi = A\Pi$$

La loi limite (s, i, m) est donc forcément l'état stable (vers lequel le système converge au bout d'un temps long).
Dans ce cas, on remarque que quelle que soit la position initiale $x0$ de la chaîne de Markov, sa loi ses stabilise vers la même loi limite.

Remarques

1. La réciproque est fausse : il peut y avoir un état stable mais les lois de la chaîne de Markov peuvent ne pas converger vers l'état stable.

2. On sait que l'état stable est un vecteur stochastique propre de A associé à la valeur propre 1. On peut donc le trouver en Scilab grâce à l'instruction `[P,D]=spec(A)` qui renvoit dans D une matrice diagonale dont les coefficients diagonaux sont les valeurs propres et P la matrice dont les colonnes forment une base de vecteurs propres associés.
   On cherche alors dans quelle colonne $i0$ de P se trouve un vecteur propre associé à la valeur propre 1.
   S'il n'est pas de signe constant, il n'y a pas de vecteur stochastique dans $E_1(A) = Vect(P)$, sinon, il y en a un donné par `pi=P(:,i0)./sum(P(i,:))`.
   Le vérifier en cherchant l'état stable de la matrice A.

2. Le Google Page Rank

Nous allons illustrer la méthode utilisée par Google pour classer les pages web par "indice de popularité".
L'idée est d'étudier l'évolution des déplacements d'un individu sur le net, et de regarder la probabilité p_i d'être sur un site i au bout d'une très grand nombre de déplacements (limite de la chaîne de Markov). p_i est alors appelée "indice de popularité de la page i".

Nous commencerons par regarder un cas très simple (et très édulcoré !) où il n'existerait sur le Web que 3 pages internet ! puis on étudiera le cas général.

a) Le cas particulier

Exercice 2.
A l'issue de sa requête, un internaute est susceptible d'aller sur trois sites A, B, et C. On suppose que l'internaute se déplace forcément vers un lien de la page Web où il se trouve.
- le site A contient un lien vers lui-même, un lien vers B, et un lien vers C ;
- le site B contient 5 liens vers lui-même, un vers A et un vers C ;
- le site C comporte un lien vers A, 7 liens vers B et 4 vers C.

A l'instant 0, l'internaute choisit au hasard l'un des trois sites. Par la suite, la probabilité de passer d'un site (à l'instant n) vers un autre (à l'instant $n+1$) est proportionnelle au nombre de liens du premier site vers le deuxième.

Pour $n \in \mathbb{N}^*$, on désigne par a_n, b_n et c_n les probabilités des évènements A_n, B_n et C_n définis par : A_n : "à l'instant n l'internaute est sur le site A", B_n : "à l'instant n l'internaute est sur le site B" et C_n : "à l'instant n l'internaute est sur le site C".

1. Représenter le modèle par un graphe (diagramme de transition).
2. En justifiant rigoureusement au moins l'une des égalités, exprimer a_{n+1}, respectivement b_{n+1} et c_{n+1}, en fonction des trois réels a_n, b_n et c_n.
3. Ecrire la matrice de transition S.
4. Mémoriser la matrice S dans Scilab puis compléter et exécuter la fonction suivante qui affiche la chaine de Markov $(X_0, X_1, ..., X_n)$ des déplacements d'un internaute :

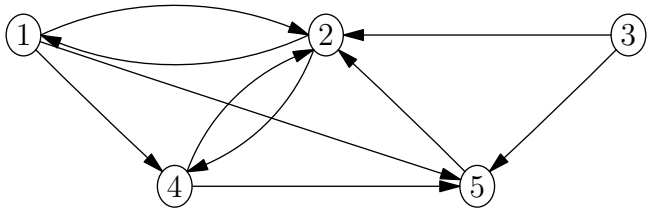
```
function X=deplacements(n)
    x0=grand(1,1,****,****,****)
    X=grand(n,'markov',S',x0)
    X= [x0 X]
endfunction
```

   Faire appel plusieurs fois à cette fonction pour un grand nombre de déplacements ($n = 100$ par exemple). Est-ce qu'un site à l'air plus visité que les autres ? Moins visité que les autres ? comment l'expliquer ?
5. Calculer avec Scilab les vecteurs de lois U_0, U_1, ..., U_{10}. On créera une matrice E de taille 3×11 dont la ième colonne contiendra le vecteur U_{i-1}.
6. Tracer sur un même graphique les courbes des suites (a_n) , (b_n) et (c_n) . Que remarque-t-on ?
   On appelle "indice de notoriété" d'une page A la valeur $a = \lim a_n$. Donner une valeur approchée des indices de notoriété a, b et c et classer les pages par indices de notoriétés.
7. Regarder la valeur exacte des indices de notoriété en calculant la valeur exacte de l'état stable (on utilisera la commande `spec` comme indiqué dans la remarque de l'exercice précédent).

b) Le cas général

Exercice 3.
On modélise l'ensemble des pages du WEB par un ensemble fini d'états $E = \{1, 2, ..., N\}$.
(le nombre N est bien très grand, de l'ordre de 10^{13}, mais reste fini).
On modélise les déplacements de l'internaute par une chaînes de Markov (X_n) où la variable X_n est le numéro de la page où se situe l'internaute après n déplacements.
Après avoir fait des études statistiques des comportements sur le Web, Google part du principe que :
• A l'instant 0, l'internaute choisit une page internet au hasard parmi les N pages.
• lorsqu'un internaute est sur une page i :
- dans 85% des cas, il se déplacera au hasard vers l'une des l_i pages pointées par i (ici, une page a au plus un lien vers une autre).
- dans 15% des cas, il se déplacera au hasard vers n'importe laquelle des pages du Web.
La matrice de transition est alors définie par : $P(j,i) = \begin{cases} 0,85 \times \frac{1}{l_i} + 0,15 \times \frac{1}{N} & \text{si i pointe vers j} \\ 0,15 \times \frac{1}{N} & \text{sinon} \end{cases}$.

Nous allons illustrer l'algorithme de Page Rank sur un groupe autonome de $N = 5$ pages différentes, ayant des liens pointant les pages les unes vers les autres, selon le diagramme suivant :



1. Taper dans la console l'instruction :

   ```
   A=0.15*1/5*ones(5,5)+0.85*[0 1/2 0 0 0; 1/3 0 1/2 1/2 0;0 0 0 0 0;1/3 1/2 0 0 0;1/3 0 1/2 1/2 1]
   ```

   et expliquer pourquoi cette matrice est bien la matrice de transition de la chaîne de Markov (on expliquera 3 coefficients) de la matrice.
2. Déterminer la valeur exacte de la loi stationnaire à l'aide de la commande `spec`.
3. *La chaine de Markov (X_n) converge-t-elle vers la loi stationnaire (état stable) ?*
   Calculer les lois des variables X_0, X_1, X_2, ..., X_{10} et les ranger dans une matrice E dont la ième colonne est la loi de X_{i-1}. Afficher E. Que remarque-t-on ?
4. Vérification graphique : afficher graphiquement les courbes des suites $(P(X_n = 1))$, $(P(X_n = 2))$, ...,$(P(X_n = 5))$.
5. Ranger les pages internet par indice de popularité. Les résultats vous semblent-ils cohérents ? Commenter.

3. Autres chaînes de Markov

Exercice 4.
Un joueur joue à un jeu au casino. On suppose qu'il dispose d'une fortune de a euros et que celle du casino est de b euros. A chaque tour, il gagne un euro avec probabilité $p \in]0, 1[$ ou perd un euro (donné au casino) avec probabilité $q = 1 - p$. Le jeu s'arrête lorsque le joueur ou le casino se retrouve ruiné.

On peut alors modéliser la situation par une chaîne de Markov homogène (X_n) où X_n représente la fortune du joueur à l'issue du nième tour.

1. représenter le modèle par un graphe (diagramme de transition) et donner la matrice de transition P. *On remarque qu'on ne peut plus quitter les états* 0 *et* $a + b$ *une fois qu'on est a atteint : ces états sont appelés* **états absorbants**.
2. Compléter et exécuter la fonction suivante qui prend en entrée a, b et p et renvoit en sortie la matrice de transition P :

   ```
   function P=transition(a,b,p)
       P=zeros(a+b+1,a+b+1)
       P(1,1)=****
   ```

```
    P(a+b+1,a+b+1)=****
    for i=2:(a+b) do
        P(i-1,i)=****
        P(i+1,i)=****
    end
endfunction
```

3. Expliquer ce que fait la fonction suivante (on expliquera précisément le rôle de chacune des instructions) puis l'exécuter :

```
function []=ruinejoueur(p,a,b,n)
    P=transition(a,b,p)
    X=grand(n,'markov',P',a+1)
    X=X-1
    X=[a X]
    plot2d([0:n],X)
endfunction
```

Remarque utile : Scilab travaille avec les numéros des états : l'état numéro 1 pour Scilab est donc l'état 0, ...l'état numéro $a+b+1$ de Scilab est donc l'état $a+b$. Il faut donc dire à Scilab que l'état initial est l'état numéro $a+1$.

4. Tester ce programme plusieurs fois dans le cas où le joueur dispose de 20 euros au départ, le casino de 30 euros et que le joueur à 2 chances sur 5 de gagner une partie (on fera un très grand nombre n de parties). Quel destin attend le joueur ?

5. Tester plusieurs fois le programme pour $p = 1/3$, $a = 20$, $b = 300$, quel destin attend le joueur ?
Recommencer pour $p = 1/3$, $a = 20$, $b = 300$, quel destin attend le joueur ?
Mêmes questions pour $p = 0.5$, $a = 20$, $b = 300$.

6. Compléter et exécuter la fonction suivante qui simule 1000 chaines de Markov $(X_0, X_1, ..., X_n)$ indépendantes et calcule à partir de ces 1000 simulations l'effectif N puis la fréquence f de l'évènement "le joueur est ruiné avant le n-ième jeux" :

```
function f=ruinejoueurfrequence(p,a,b,n)
    P=transition(a,b,p)
    N=0
    for k=1:1000 do
        X=grand(n,'markov',P',a+1)
        X=X-1
        if ************* then
            N=N+1
        end
    end
    f=************
endfunction
```

Tester ce programme pour $a = 20$, $b = 30$ et différentes valeurs de p ($p < 1/2$, $p \geq 1/2$, p égal ou très proche de $1/2$) et estimer la probabilité de l'évènement "le joueur est ruiné" dans chacun des cas.
La valeur obtenue est-elle la valeur exacte de la probabilité que le joueur soit ruiné ? Pourquoi ? Quel est le lien entre fréquence et probabilité ?

Exercice 5.
On veut modéliser la mobilité sociale d'une population. Pour cela, on s'intéresse à la profession des individus d'une génération, comparée avec celle de leurs parents.

L'INSEE fournit par exemple le tableau suivant, à partir d'un sondage effectué sur 18373 individus de sexe masculin :

	Agr.	Indép.	Cadres	Interm.	Emp.	Ouvr.
Agr.	621	24	10	9	7	43
Indép.	237	550	127	138	210	621
Cadres	220	479	749	552	490	700
Interm.	405	439	325	568	643	1520
Emp.	214	183	100	182	393	855
Ouvr.	1126	499	116	352	770	3896
Total	2823	2174	1427	1801	2513	7635

où on a utilisé les abréviations : Agr.=Agriculteur, Indép.=Indépendants,Iterm.=Professions intermédiaires, Emp.=Employés et Ouvr.=Ouvriers.

L'effectif placé ligne i colonne j représente le nombre de sondés qui exerce la profession i, sachant que leur père exerçait la profession j.
Une colonne donne donc les effectifs des catégories professionnelles des sondés, pour une catégorie socio-professionnelle fixée pour le père : on visualise ainsi sur chaque colonne les "transitions intergénérationnelles".

On modélise la situation par une chaîne de Markov (X_n) à 6 états : X_n donne la catégorie socio-professionnelle de l'individu de la n-ième génération.

1. Taper les instructions suivantes :

```
B=[612 24 10 9 7 43;237 550 127 138 210 621;220 479 749 552 490 700;405 439 325 568 643 1520; 21
A=zeros(6,6)
for i=1:6 do
    A(:,i)= B(:,i)/sum(B(:,i))
end
```

   Expliquer pourquoi la matrice A ainsi créée est bien la matrice de transition de cette chaîne de Markov.

2. Ecrire une `function Y=socio(x0,n,r)` qui simule r fois la suite de n variables $(X_1, X_2, ..., X_n)$, partant de l'état $x0$, et qui donne en sortie un vecteur Y contenant les r valeurs de X_n obtenues.
   Cette fonction nous donne donc la profession au bout de n générations de r individus provenant tous à l'instant 0 de la même catégorie $x0$.
   En exécutant la fonction `socio` avec des paramètres bien choisis, estimer la probabilité que l'arrière petit fils d'un ouvrier soit cadre.
   On fera cette estimation sur 100 simulations de l'expérience.

3. Modifier la fonction `socio` pour qu'elle construise le diagramme empirique de la loi de X_n, obtenu à l'aide de r simulations partant de la catégorie socioprofessionnelle x_0 *(cette fonction ne renverra plus de sortie).*

4. Taper à la suite et exécuter les instructions suivantes :

```
for i=1:6 do
    scf(i)
    socio(i,100,1000)
end
```

   Que représentent les différents diagrammes obtenus ?
   Les résultats sont-ils cohérents avec la théorie sur les chaînes de Markov ? Le vérifier en calculant la valeur exacte de l'état stable.
   Quelle(s) conclusion(s) un sociologue pourrait-il tirer de ces résultats ?
   Quelles critiques pourriez-vous faire de ces conclusions ?