Outline

Modélisation stochastique, Simulation

Stephan Robert, HEIG-Vd

8 octobre 2009

Introduction

Etapes pour la simulation

Avec la simulation nous essayons de **reproduire le comportement d'un système**. Ensuite nous essayons d'estimer certaines caractéristiques (performance par exemple).

- **Modélisation** du système comme un processus stochastique dynamique
- Génération de différentes (plusieurs) **réalisations** du processus stochastique à l'aide de générateurs de nombres aléatoires.
- **Mesures** des résultats obtenus
- **Analyse** statistique et conclusions

Alternatives à la simulation (2)

Chapitre précédent : **Analyse mathématique**

- Modéliser le système comme un processus stochastique (modes de transmission par exemple)
- Résolution du modèle à l'aide d'outils mathématiques

La **phase de modélisation** est présente pour les deux approches.

La **précision** peut être différente pour les deux méthodes...

Exemple : Date d'anniversaire

Combien faut-il réunir de personnes pour que la probabilité que deux personnes aient leur anniversaire le même jour soit d'au moins 0.5 ? Les gens sont numérotées de 1 à n.

- Si $n = 1$ alors il y a 365 possibilités de dates d'anniversaire, équiprobables (admis).
- Si $n = 2$ il y a 365^2 possibilités de dates d'anniversaire.

Observation : Lorsque $n = 2$, il y a 365 possibilités que les dates d'anniversaire tombent sur le même jour. Donc il a y $365^2 - 365$ possibilités pour que les dates d'anniversaire ne tombent pas en même temps, ou $365 \times (365 - 1) = 365 \times 364$.

Exemple : Date d'anniversaire (2)

En généralisant : Probabilité que aucun anniversaire ne tombe en même temps parmi n personnes :

$$\frac{365 \times 364 \times 363 \times \ldots \times (365 - n + 1)}{365^n}$$

Autrement dit la probabilité qu'au moins deux personnes aient leurs anniversaires en même temps (p_n) est de

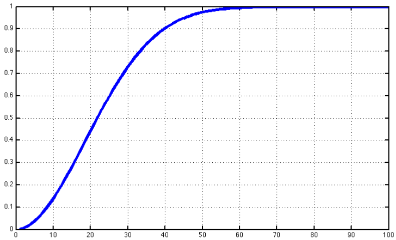
$$p_n = 1 - \frac{365 \times 364 \times 363 \times \ldots \times (365 - n + 1)}{365^n}$$

Exemple : Date d'anniversaire, exemple numérique

Exemple numérique :

- $n = 23$, $p_n = 0.5073$
- $n = 40$, $p_n = 0.8912$
- $n = 75$, $p_n = 0.9997$

Question : Comment résoudre le problème avec une simulation ?



Avantages-Inconvénients de la simulation

Avantages

- Hypothèses moins restrictives
- Le modèle et les paramètres peuvent être changés rapidement.
- Utilisation possible de mesures réelles dans les simulations.
- Certains résultats ne peuvent être obtenus que par simulation

Désavantages

- Prend beaucoup de temps de calcul
- La validation prend du temps
- Les résultats peuvent être difficiles à interpréter, les relations entre variables difficile à établir.
- Les résultats peuvent être inexacts, surtout si le temps de simulation est trop court ou si l'analyse est trop courte.
- Les simulations peuvent être complexes.

Avantages-Inconvénients des calculs analytiques

Avantages

- Les résultats sont obtenus rapidement.
- Les résultats du modèle sont exacts.
- Donnent de bonnes indications sur le comportement du modèle
- Optimisation possible (souvent difficile)
- Analyse souvent rapide.

Désavantages

- Hypothèses souvent trop restrictives (par exemple : stationarité).
- Les calculs-preuves sont souvent très difficiles, voire impossible.

Classification des types de simulations

Classification générale

- Temps discret, états discrets.
- Temps discret, états continus.
- Temps continu, états discrets.
- Temps continu, états continus.

Classification temporelle

- **Temps continu** : Le temps de l'ordinateur est toujours discret. On prend des intervalles aussi petits que possible pour observer tous les changements d'états.
- **Temps discret** : On observe le système à des temps déterminés : $t, t + \Delta t, \ldots$.
- Simulations à **événements discrets** : Focalisation sur le système à chaque fois qu'il se “passe” quelque chose. On a une **liste d'événements**.

Programmes de simulation, implémentation

1. Programme généraux
   - C/C++ : Fonctionnent pour tout type de simulations, très flexibles, sans fonctions spécifiques pour la simulation.
   - Matlab, Octave, ... : Principal inconvénient : Lenteur.
2. Programmes spécialisés dans le domaine des communications
   - Opnet : Réseaux en général
   - *ns-2* : Fiable, standard de facto dans le monde de la recherche.

Génération de nombres pseudo-aléatoires

Réalisations de variables aléatoires

- Les réalisations sont basées sur un **générateur** de nombres pseudo-aléatoires.
- Premier pas :
  - Génération de nombres **pseudo-aléatoires "uniformément" distribués** dans l'intervalle $[0, 1]$.
- Deuxième pas : Passer de la distribution uniforme à la distribution désirée.
  - **Discrétization** (loi de Bernoulli, binômiale, Poisson, Géométrique)
  - Une possibilité : **transformation inverse**
  - Sinon : autres méthodes (pour la loi normale) ou **méthode d'acceptation et de rejet** (où deux variables aléatoires uniformément distribuées dans l'intervalle $[0, 1]$ doivent être générées.

Générateur de nombres aléatoires

- Un **générateur de nombres aléatoires** est un algorithme qui génère des nombres **pseudo**-aléatoires $\Rightarrow$ Générateur de nombres pseudo-aléatoires ! ! ! Génération de nombres entiers Z_i dans l'intervalle $0, 1, 2, 3, \ldots, m-1$.
- Remarques :
  - La séquence générée est *périodique* (avec pour but d'avoir la période la plus grande possible !)
  - Les nombres générés ne sont pas aléatoires du tout ! mais complétement déterministes, d'où le nom *"pseudo"*.
  - La séquence générée est *déguisée* : On ne remarque pas qu'elle est déterministe ! Les nombres produits apparaîssent comme étant indépendants et uniformément distribués dans $\{0, 1, 2, 3, \ldots, m-1\}$
- Validation : *Tests statistiques* pour vérifier l'indépendance des nombres générés et l'uniformité de la distribution.

Générateur de nombres aléatoires (2)

- Caractéristiques souhaitées :
  - Rapide
  - Peu complexe
  - Portable
  - Longs cycles
  - Les nombres doivent être indépendants
  - Les nombres doivent avoir une distribution définie

Générateur congruentiel multiplicatif

- **Le générateur congruentiel multiplicatif** utilise l'algorithme suivant pour générer des nombres dans $\{0, 1, 2, 3, \ldots, m-1\}$:

$$Z_n = aZ_{n-1} \bmod m$$

- Paramètres :
  - $Z_0 < m$: "graine".
  - a, m : entiers positifs. Doivent être choisis avec soin !
  - m : Nombre premier. Période : $m-1$.
  - a : Racine primitive modulo m.
  - Exemple (Cray Research) : $m = 2^{48}$ et $a = 44485709377909$.

Générateur congruentiel multiplicatif (2)

Calcul de $19 * 7$ modulo 23 :

- Calcul en arithmétique normale, par exemple $19 * 7 = 133$
- Diviser 133 par 23 : 5 $(5 * 23 = 115)$, reste : $133 - 115 = 18$, donc $19 * 7 \bmod 23 = 18$.

Exemple :

- $Z_0 = 7$, $a = 19$ et $m = 23$
- $Z_1 = 19 * 7 \bmod 23 = 133 \bmod 23 = (5 * 23 + 18) \bmod 23 = 18$,
- $Z_2 = 19 * 18 \bmod 23 = 342 \bmod 23 = (322 + 20) \bmod 23 = 20$, etc...

Générateur congruentiel linéaire

- **Le générateur congruentiel linéaire** utilise l'algorithme suivant pour générer des nombres dans $\{0, 1, 2, 3, \ldots, m-1\}$:

$$Z_n = (aZ_{n-1} + c) \bmod m$$

- Paramètres :
  - $Z_0 < m$: "graine".
  - a, m, c : entiers positifs. Doivent être choisis avec soin !
  - m et c sont premiers entre eux. Période maximale : $m - 1$.
  - $a - 1$ est divisible par tous les facteurs premiers de m et est un multiple de 4 si m est un multiple de 4.

Générateurs pseudo-aléatoires utilisés en pratique

- Valeurs utilisées en pratique

Source	m	a	c	x_0
ANSI C, BSD : rand()	2^{31}	1103515245	12345	12345
Apple	$2^{35}-1$	5^{13}	0	1
NAG	$2^{59}-1$	13^{13}	0	$2^{32}+1$
Cray	2^{48}	44485709377909	0	1
Maple	$10^{12}-11$	427419669081	0	1

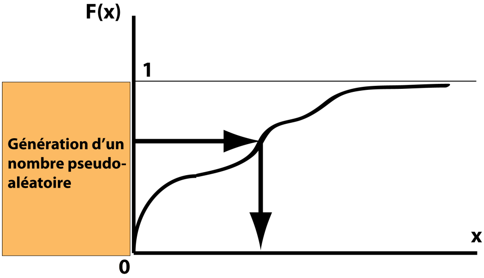
- Matlab (2007), Mapple : *Mersene Twister*

- Remarque : Soit Z_n le nombre généré dans l'intervalle $\{0, 1, 2, 3, \ldots, m-1\}$, alors Z_n/m est à peu près "distribué uniformément" dans l'intervalle $[0, 1]$ (selon les tests statistiques).

Génération de réalisations de variables aléatoires

Méthode de **transformation inverse**

Variable aléatoire discrète : $F(x_n) = \sum_{i=1}^{n} P(X = x_i)$



Méthode de **transformation inverse** : Algorithme pour une fonction de répartition arbitraire.

Si $U < p(x_0)$ alors on fixe $X = x_0$
Si $p(x_0) \leq U < p(x_0) + p(x_1)$ alors on fixe $X = x_1$
Si $p(x_0) + p(x_2) \leq U < p(x_0) + p(x_1) + p(x_2)$ alors on fixe $X = x_2$
$\ldots$
Si $\sum_{i=0}^{n-1} p(x_i) \leq U < \sum_{i=0}^{n} p(x_i)$ alors on fixe $X = x_n$

Simulation de variables aléatoires exponentielles

Fonction de répartition pour la loi exponentielle :

$$F(x) = 1 - e^{-\lambda x}$$

En développant :

$$x = -\frac{1}{\lambda} \ln(1 - F(x))$$

En remplaçant $1 - F(x)$ par U (généré par le générateur de nombres aléatoires, $= 1 - U$) :

$$\boxed{x = -\tfrac{1}{\lambda} \ln(U)}$$

Simulation de variables aléatoires gaussiennes

Méthode de **Box-Miller** : Les deux variables aléatoires ci-dessous sont iid et normalement distribuées :

$$X_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2) \sim \mathcal{N}(0,1)$$
$$X_2 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2) \sim \mathcal{N}(0,1)$$

avec U_1, U_2 générés dans $[0,1]$.

Pour obtenir une distribution de $Y \sim \mathcal{N}(\mu, \sigma^2)$ il suffit de savoir que $Y = \mu + \sigma X$ avec $X \sim \mathcal{N}(0,1)$.

Estimation de paramètres

(Introduction)

estimation de la moyenne et de la variance

Hypothèse : Les variables aléatoires X_i sont iid, d'espérance mathématique μ et de variance σ^2, alors

- **Estimation de la moyenne**

$$\hat{\mu}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n-1}{n}\left(\bar{X}_{n-1} + \frac{X_n}{n-1}\right)$$

- **Estimation de la variance** (non biaisé)

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 =$$
$$\frac{n-2}{n-1}\left(Var(X_{n-1}) + \frac{(X_n-\bar{X})^2}{n-2}\right)$$

Convergence et propriétés des estimateurs : On y reviendra !

Intervalles de confiance

(Rappels)

Intervalles de confiance : Généralités

Observation : l'estimation n'est jamais exacte. On construit donc un **intervalle de confiance** dans lequel l'estimation doit se trouver avec une probabilité donnée (typiquement 0.9, 0.95, 0.99).

Principe : Considérons un estimateur (variable aléatoire !) T du paramètre θ qui a une certaine densité de distribution $f_T(\theta)$. Alors nous pouvons construire un intervalle de confiance :

- On fixe un **risque d'erreur** α, ou **seuil de confiance** $1 - \alpha$ de telle sorte à ce que

$$P(|T - \theta| \leq c) = 1 - \alpha$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue

Faisons l'hypothèse que les variables aléatoires X_i, $i = 1, \ldots, n$ sont iid avec une **espérance mathématique inconnue,** μ et une **variance connue,** σ^2.

Nous savons par le **théorême central limite**, pour n grand, que

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1)$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue (2)

Remplaçons T par $\bar{X}_n$ et θ par μ :

$$P(|\bar{X}_n - \mu| \leq c) = 1 - \alpha$$

Ce qui est équivalent à

$$P(\frac{|\bar{X}_n - \mu|}{\sigma/\sqrt{n}} \leq \frac{c}{\sigma/\sqrt{n}}) = 1 - \alpha$$

$$P(-\frac{c}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{c}{\sigma/\sqrt{n}}) = 1 - \alpha$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue (3)

Soit $F(x)$ la fonction de répartition de la loi normale réduite :

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

En reprenant les calculs :

$$F\left(\frac{c}{\sigma/\sqrt{n}}\right) - F\left(-\frac{c}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

En remarquant que $F(-x) = 1 - F(x)$, il vient

$$F\left(\frac{c}{\sigma/\sqrt{n}}\right) - \left(1 - F\left(-\frac{c}{\sigma/\sqrt{n}}\right)\right) = 1 - \alpha$$

$$F\left(\frac{c}{\sigma/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue (4)

$$F\left(\frac{c}{\sigma/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}$$

$$\frac{c}{\sigma/\sqrt{n}} = F^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Si Z représente la variable aléatoire normale réduite,

$$\frac{c}{\sigma/\sqrt{n}} = z_{1-\frac{\alpha}{2}}$$

ou

$$c = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue (5)

et donc

$$P(|T-\theta| \leq z_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Proposition : L'intervalle de confiance pour l'estimation de la moyenne avec un niveau de confiance de $(1-\alpha)$ est donné par

$$\bar{X} \pm z_{1-\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$$

Intervalles de confiance : Espérance mathématique inconnue, variance connue (6)

seuil de confiance	α	$\alpha/2$	$1-\alpha/2$	$z_{1-\alpha/2}$
80%	0.2	0.1	0.9	1.28
90%	0.1	0.05	0.95	1.645
95%	0.05	0.025	0.975	1.960
96%	0.04	0.02	0.98	2.05
98%	0.02	0.01	0.99	2.33
99%	0.01	0.005	0.995	2.58
99.9%	0.001	0.0005	0.9995	3.291

Intervalles de confiance : Espérance mathématique inconnue, variance inconnue

Souvent nous ne connaissons **ni l'espérance mathématique, ni la variance**, qui peut cependant être estimée avec l'estimateur de la variance, alors

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \rightarrow \text{Student(n-1)}$$

où (Student(n-1)) est une loi de Student à (n-1) degrés de liberté (tend vers la **loi normale** que n devient grand !).

Intervalles de confiance : Espérance mathématique inconnue, variance inconnue (2)

Avec un raisonnement identique au précédent, nous pouvons construire un intervalle de confiance pour estimer μ :

Proposition : L'intervalle de confiance pour l'estimation de la moyenne avec un niveau de confiance de $(1-\alpha)$ est donné par

$$\bar{X} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

où les valeurs de $t_{n-1,1-\frac{\alpha}{2}}$ sont données par la loi de Student.

Remarque : Pour $n \geq 30$, $t_{n-1,1-\frac{\alpha}{2}}$ sera remplacé par $z_{1-\alpha/2}$.

Table de Student (Wikipédia)

n	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385

Observations

- Les résultats sont plus précis (l'intervalle de confiance se réduit) lorsque
  - le nombre de simulations n augmente
  - la variance de chaque simulation est réduite, soit en laissant la simulation "tourner" plus longtemps ou en utilisant des techniques de réduction de variance (à voir dans le chapitre "estimation").
- Etant donné une précision voulue pour les résultats, il est possible de déterminer le nombre de simulations qu'il faut effectuer.
- Suite au prochain épisode :-)…. **We'll be back !**