La simulation en statistique (suite)

Bernard Egger

Résumons-nous : pour « construire le hasard », il est courant d'employer des suites récurrentes « chaotiques ». De telles suites, phénomènes purement déterministes comme toute suite récurrente, ont l'étrange propriété de ressembler terriblement à ce que l'on attend du hasard. Cette similitude nous intéresse d'autant plus qu'elle n'est pas seulement d'ordre graphique. Il est par exemple courant de se demander que la variable aléatoire qui à un nombre obtenu par cette suite associe le chiffre d'un rang donné dans ce nombre suit bien une loi uniforme. On pourrait tout aussi bien vouloir construire une suite qui permettrait d'avoir une loi binomiale, une loi de Poisson, ou une loi normale.

Dans tous les cas, nous construisons des générateurs de nombres aléatoires ou plutôt pseudo-aléatoires, qu'il est important de tester, pour savoir s'ils sont conformes à notre attente. Indépendamment des tests, nous avons vu dans l'article précédent qu'un grain de sable pouvait surgir à l'insu de l'utilisateur : la cyclicité du générateur. Le problème est d'importance : un test peut se révéler adapté si on l'effectue sur de petits échantillons, et totalement inadapté si on le fait fonctionner sur un échantillon important : nous avons vu que la suite proposée dans le magazine 3'33 était périodique avec une période de 100000. Nous voici bien loin du hasard ! Il n'y a d'ailleurs dans le cadre des suites chaotiques aucune solution évidente : votre calculatrice, ou le logiciel Excel utilisent des nombres décimaux d'une longueur maximale donnée. Quel que soit celui dont on part, il arrivera un moment où nous aurons épuisé tous les nombres possibles, et la suite étant récurrente, totalement déterministe, un nouveau cycle démarrera. La question principale est alors de construire des générateurs dont le cycle est le plus long possible. 100000 c'est vraiment court !

Nous aborderons dans un prochain article les réponses qui sont le plus fréquemment apportées à cette question. Pour le moment revenons aux tests.

La dernière fois nous avons examiné l'uniformité. Elle constitue souvent le minimum exigible. De façon plus élaborée on peut tester des situations faisant appel à l'équiprobabilité mais moins directement que dans la répartition uniforme. Le principe général en est le suivant : retrouver des résultats théoriques connus (en tenant compte de l'imprécision produite nécessairement par la fluctuation d'échantillonnage). Bien entendu dans un second temps, les mêmes procédés pourront être utilisés pour mettre les élèves sur la voie de ces probabilités.

Un exemple simple est constitué par une situation de Bernoulli. Prenons un cas classique : un questionnaire comprend 9 questions, comportant chacune trois réponses possibles, une seule étant exacte. On répond au hasard. On sait que la variable aléatoire X qui donne le nombre de réponses justes suit une loi binomiale de paramètres 9 et $\frac{1}{3}$. Son espérance mathématique vaut donc 3. 1000 candidats vont répondre aléatoirement au questionnaire. On se propose de simuler cette situation et de « vérifier expérimentalement » que la moyenne des réponses justes est proche de 3.

Les générateurs de nombres pseudo-aléatoires des calculatrices ou des tableurs retournent habituellement des nombres compris entre 0 et 1. Il y a trois réponses possibles ; deux sont fausses. On notera 0 la réponse juste et 1 une réponse fausse. Pour chaque question, il faut obtenir une simulation (c'est-à-dire une variable aléatoire) qui retourne 0 ou 1 avec « deux fois plus de 1 que de 0 ». Si x est un nombre compris entre 0 et 1, $\text{PARTIE\_ENTIÈRE}(3x)$ est soit 0, soit 1, soit 2. Soit y l'un de ces trois entiers, alors $\text{PARTIE\_ENTIÈRE}(y/2+0,5)$ vaut 0 quand y vaut 0 et vaut 1 si y vaut 1 ou 2. On compose tout cela et l'on prend $x = \text{ALEA}()$ (nombre aléatoire). On a comme instruction :

$$\text{PARTIE\_ENTIÈRE}(\text{PARTIE\_ENTIÈRE}(3 * \text{ALEA}()))/2 + 0,5)$$

A1 = =ENT(ENT(3*ALEA())/2+0,5)

A	B	C	D	E	F	G	H	I
0	1	0	1	1	0	1	1	1

Il ne reste plus qu'à « compter » le nombre de 0, ou plutôt qu'à apprendre au tableur à compter. L'instruction « NB.SI » permet de compter le nombre d'occurrences de *variable* dans la plage : *plage* (NB.SI (plage : variable)).

Par exemple avec ces nouvelles données :

J1	=	=NB.SI(A1:I1;0)							
A	B	C	D	E	F	G	H	I	J
1	0	0	1	1	1	1	1	1	2

Il ne reste plus qu'à recopier cette première ligne jusqu'à la millième ligne

990	1	1	0	1	1	0	1	0	0	4
991	1	1	1	1	1	1	1	1	1	0
992	1	1	0	0	0	0	1	1	1	4
993	0	1	0	1	1	1	1	1	1	2
994	0	1	1	1	1	1	0	1	0	3
995	1	0	0	1	0	0	1	0	1	5
996	1	1	0	1	1	0	1	0	0	4
997	1	1	0	0	1	0	0	0	1	5
998	1	1	1	1	1	1	1	1	1	0
999	1	1	1	1	0	1	1	1	1	1
1000	1	1	0	1	1	1	1	1	1	1

On peut alors calculer la moyenne de la colonne J.

=	=MOYENNE(J1:J1000)		
D	E	F	G
3,032			

Ce résultat paraît satisfaisant. Son utilisation pédagogique aussi ! L'espérance est proche de 3 ce qui rend « facilement » identifiable à $\frac{1}{3}$.

D'un point de vue plus théorique, on peut se demander si ce résultat est « suffisamment » proche de 3. Le nombre de réponses justes à chaque questionnaire est une variable aléatoire de moyenne 3 et de variance 2.

Le théorème de la limite centrée permet de dire alors que la variable aléatoire qui à chaque échantillon de 1000 questionnaires associe le nombre moyen de réponses justes par questionnaire suit une loi normale de moyenne 3 et d'écart-type $\frac{\sqrt{2}}{\sqrt{1000}}$. 95 % des échantillons de 1000 questionnaires auront donc un nombre moyen de réponses justes compris entre $3 - 1,96 \times \frac{\sqrt{2}}{\sqrt{1000}}$ et $3 + 1,96 \times \frac{\sqrt{2}}{\sqrt{1000}}$, c'est-à-dire environ entre $2,91$ et $3,09$. Le résultat obtenu de $3,032$ est conforme à cette attente. Le générateur de nombres pseudo-aléatoires d'Excel « résiste » bien à ce test.

Mais on peut aller plus loin. Il est en effet très facile de déterminer combien chacune des valeurs entre 0 et 9 apparaît dans la plage des 1000 résultats. On en déduit aisément la fréquence de chacune d'elles.

Or nous connaissons (les élèves pas encore si l'on utilise l'activité comme découverte), les résultats théoriques fournis par la loi binomiale. Voilà une bonne occasion d'utiliser un test du Khi-deux. Ce test ne s'utilisant qu'avec des effectifs, nous ramenons les fréquences et autres probabilités à des effectifs dont le somme vaut 1000.

	A	B	C	D	E	F	G	H	I	J
1	0	1	2	3	4	5	6	7	8	9
2	26	135	229	257	208	100	32	12	1	0
3	26	117,1	234,1	273,1	204,8	102,4	34,1	7,3	0,9	0,01

La deuxième ligne correspond aux effectifs obtenus sur l'échantillon et la troisième aux probabilités théoriques multipliées par 1000.

Autre problème : les deux dernières classes ont des effectifs théoriques trop petits (inférieurs à 5). Leur poids relatif par rapport à l'ensemble étant faible, on appliquera le test sur les 7 premières colonnes (c'est-à-dire sur un effectif total de 999).

On obtient alors comme résultat environ : $0,42$. Dans 58 % des cas on obtient des échantillons « aussi bons voire meilleurs » c'est-à-dire plus conforme à une loi binomiale, mais en tenant compte de la fluctuation d'échantillonnage, au seuil de confiance de 95 %, notre échantillon ne permet pas de rejeter l'hypothèse selon laquelle on peut ajuster la distribution obtenu à chaque questionnaire par une loi binomiale.

En définitive, on obtient pour nos deux tests une « confirmation » indirecte de la valeur de la distribution uniforme des nombres pseudo-aléatoires fournis par Excel.

Après la loi binomiale, la loi normale est assez simple à tester. Pour simuler une loi normale, on utilise la méthode dite de Box-Muller : si X_1 et X_2 sont deux variables aléatoires uniformes sur $]0;1]$, alors la variable $Y = \sqrt{-2\ln X_1}\cos(2\pi X_2)$ suit une loi normale centrée réduite.

On suppose que le générateur d'Excel correspond bien à une loi uniforme sur $]0;1]$. On place par exemple 5000 nombres par la fonction ALEA () sur les colonnes A et B. Ces deux colonnes correspondront aux variables X_1 et X_2. On construit la variable Y sur la colonne C.

	A	B	C	D	E	F	G	H	I	J
1	0,8314938	0,31193083	-0,23047266	-10	0		0			
2	0,96758474	0,86107526	0,16497046	-3	12	12	0,0013	6,7498		
3	0,60055278	0,24017475	0,062304	-2,7	22	10	0,0035	10,585		
4	0,05930925	0,68744417	-0,91040479	-2,4	45	23	0,0082	23,653		
5	0,7380014	0,12308807	0,55777107	-2,1	90	45	0,0179	48,334		
6	0,93545164	0,28189865	-0,07272764	-1,8	171	81	0,0359	90,33		
7	0,6081676	0,02532574	0,98470135	-1,5	323	152	0,0668	154,38		0,4307809
8	0,15272506	0,12416219	1,37800649	-1,2	562	239	0,1151	241,31		
9	0,05957818	0,32867747	-1,12685822	-0,9	890	328	0,1841	344,95		
10	0,38707475	0,67242116	-0,64531127	-0,6	1314	424	0,2743	450,96		
11	0,56939818	0,27086608	-0,13874198	-0,3	1846	532	0,3821	539,18		
12	0,7169604	0,47941738	-0,80895054	0	2472	626	0,5	589,56		
13	0,59582328	0,79172239	0,26372862	0,3	3086	614	0,6179	589,56		
14	0,77341858	0,9552372	0,68868002	0,6	3659	573	0,7257	539,18		
15	0,62655934	0,65060198	-0,56540818	0,9	4109	450	0,8159	450,96		
16	0,76502373	0,50935905	-0,73064816	1,2	4418	309	0,8849	344,95		
17	0,47525855	0,72554258	-0,18670753	1,5	4670	252	0,9332	241,31		
18	0,98179797	0,44916647	-0,18198128	1,8	4827	157	0,9641	154,38		
19	0,44846112	0,547986	-1,20931145	2,1	4908	81	0,9821	90,33		
20	0,11917848	0,6631949	-1,07001624	2,4	4956	48	0,9918	48,334		
21	0,60999573	0,97942245	0,98598737	2,7	4980	24	0,9965	23,653		
22	0,76765185	0,32017093	-0,31033737	3	4989	9	0,9987	10,585		
23	0,74779808	0,33627495	-0,3933333	10	5000	11	1	6,7498		

La plupart des valeurs de la colonne C sont comprises entre -3 et 3. Dans la colonne D on fait apparaître les nombres de -3 à 3 écrites de $0,3$ en $0,3$ et deux valeurs extrêmes très différentes (-10 et 10). La colonne E comptabilise le nombre de valeurs de la colonne C inférieure à la cellule D correspondante (Par exemple ici, il y a 45 valeurs de C inférieures à $-2,4$). Pour cela on utilise encore la fonction « NB.SI ». En E1, on écrit la formule : = NB.SI (C$1 : C$2000; < & D1). Puis l'on recopie la formule jusqu'à D23. La colonne F contient le nombre de valeurs de C comprises entre deux valeurs consécutives de D. Par exemple le 23 en F4 signifie qu'il y a 23 valeurs de C comprises entre $-2,7$(D3) et $-2,4$(D4). Si l'on appelle Z une variable aléatoire qui suit la loi normale centrée réduite, on écrit en colonne G la probabilité $\mathrm{p}(Z < t)$ avec $t =$ D1 . . . D23 (= LOI.NORMALE.STANDARD (D1)). Puis en colonne H, les effectifs théoriques pour 5000 essais d'obtenir Z entre t_1 et t_2 avec $t_1 =$ D1 . . . D22 et $t_2 =$ D2 . . . D23. Par exemple en H1, on trouve la formule $= 5000 * (\mathrm{G2} - \mathrm{G1})$.

On applique alors le test du Khi-deux aux colonnes F et H. On trouve $0,43$ environ. Là encore, compte tenu de la fluctuation d'échantillonnage, on ne peut pas mettre en question que la variable Y suit bien une loi normale centrée réduite. Et donc indirectement on doit admettre que le générateur aléatoire d'Excel suit bien une loi uniforme. Ce dernier test est un peu différent des précédents puisqu'il ne prend pas une décimale particulière, mais le nombre pseudo-aléatoire dans son entier. Il existe d'autres tests qui s'intéressent à un certain nombre de décimales particulières ou à l'écart entre deux décimales identiques . . .

Dans le prochain article, nous verrons comment « réagit » le nombre π face à ces différents tests.

À suivre . . .