À PROPOS DU PROGRAMME DE STATISTIQUE EN SECONDE : REMARQUES SUR LA SIMULATION INFORMATIQUE

Michel HENRY
IREM de Franche-Comté

Une définition :

La simulation est *l'expérimentation sur un modèle*. C'est une procédure de recherche scientifique qui consiste à réaliser une reproduction artificielle (*modèle*) du phénomène que l'on désire étudier, à observer le comportement de cette reproduction lorsque l'on fait varier expérimentalement les actions que l'on peut exercer sur celle-ci, et à en induire ce qui se passerait dans la réalité sous l'influence d'actions analogues.

(Encyclopédie Universalis)

L'ordinateur est-il un véritable générateur de hasard ? En principe, pas pour le moment, tant que des phénomènes physiques, comme l'exploitation du bruit électronique par exemple, ne sont pas introduits à cette fin. Actuellement, les nombres pseudo-aléatoires que l'ordinateur (ou la calculatrice) fournit sont déterminés, dès lors qu'il a commencé leur calcul sur la base d'une initialisation (randomization) qui, en principe, détermine à chaque fois des suites différentes.

Mais la complexité de leur détermination, supposant un calcul lourd, rend impossible leur prévision par tout autre moyen humain. « *On a alors le phénomène fortuit* » selon l'expression de Poincaré. **Tout se passe donc comme si** les chiffres fournis par la fonction Random de l'ordinateur étaient issus d'un tirage aléatoire de boules numérotées de 0 à 9 dans une urne. La condition est que cette génération pseudo-aléatoire vérifie différents tests d'uniformité. En tout cas, les moyens de contrôle à notre disposition ne permettent pas d'invalider cette **hypothèse**, et nous pouvons déclarer que l'ordinateur **simule** les tirages « au hasard » successifs de ces boules de l'urne. Le problème d'obtenir un tel comportement de l'ordinateur est un problème de spécialiste : on sait que la suite « aléatoire » de ces chiffres générés est en fait périodique sur une très longue période, mais pour notre simulation, nous limitant à un nombre raisonnable de données, ce problème ne se pose pas.

Mais, par exemple, en « simulant » un sondage sur Excel, ayant précisé notre modèle probabiliste sous-jacent (loi de Bernoulli $B(1, p)$ pour la variable parente X_0 , où p est la proportion dans la population P à estimer), ayant introduit la valeur p dans l'ordinateur, avons-nous réellement simulé quelque chose ? Ou avons-nous seulement vérifié que le résultat de la simulation (conforme à ce qui était donc attendu), confirme que le fabricant de l'ordinateur et le concepteur du logiciel ont bien rempli leurs cahiers des charges ? Sans doute, en réalité c'est ce que nous avons fait.

Cependant, ne négligeons pas l'intérêt de l'outil informatique. Il permet une approche expérimentale des situations de sondages, et si la proportion p est cachée aux élèves, nous avons un outil de résolution de problèmes jouant le même rôle que les calculettes graphiques

quand elles tracent des courbes représentatives de fonctions données, ce qui suppose une bonne dose de confiance (parfois mal placée car trahie dans certaines conditions particulières) dans le bon fonctionnement de cet outil.

Mais notre question est plus fondamentalement didactique. L'ordinateur nous permet d'abord de travailler sur de vastes séries statistiques, ce qui donne aux résumés statistiques toute leur importance. Il nous permet ensuite une présentation animée du fonctionnement et de l'interaction des notions de fréquence et de probabilité, introduites auparavant, soit théoriquement, soit au travers d'expériences pratiques en nombre trop limité pour pouvoir déboucher valablement sur une bonne compréhension de la loi des grands nombres.

Même si, dans son fonctionnement réel, l'ordinateur ne fait pas intervenir de notion de probabilité. Il se limite à exhiber les effets sur les fréquences affichées du principe d'équirépartition des chiffres pseudo-aléatoires qu'il génère, principe qui entre dans ses spécifications), l'usage que nous en faisons dans la classe, **comme pseudo-générateur de hasard**, permet de faire comprendre **en actes** ces notions de fréquence, de fluctuations d'échantillonnages et de probabilité.

Mais on ne peut se satisfaire de la seule exploitation de la puissance et la rapidité de l'ordinateur pour générer des nombres aléatoires, permettant de présenter aux élèves une grande richesse de nouvelles expériences aléatoires, car cela n'aurait qu'un intérêt limité. Son intérêt didactique, **en tant qu'outil de simulation**, tient plus essentiellement en ce qu'il nous oblige à analyser la situation aléatoire en jeu et à émettre des hypothèses de modèle. Ces hypothèses portent notamment sur la loi de probabilité idoine pour représenter l'intervention du hasard dans l'expérience réelle, par exemple sur le choix de la valeur de la probabilité de Bernoulli à implanter. Il reste alors à traduire ces hypothèses en instructions informatiques pour que l'ordinateur nous permette de résoudre des problèmes éventuellement inaccessibles par le calcul *a priori*. De ce point de vue, cela suppose de comprendre le processus de modélisation et d'interpréter les résultats obtenus, rapportés aux hypothèses de modèle introduites. Simulant ainsi une expérience aléatoire réellement effectuée, au-delà de l'exploration statistique, l'ordinateur devient un outil didactique majeur pour l'apprentissage de la modélisation en probabilités.