

Entre réel et virtuel, la simulation en statistique

Nouveaux programmes, nouveaux enjeux

Bernard EGGER

Le 9 juin dernier, lors du colloque organisé pour l'épreuve finale de Mathématiques sans Frontières, j'ai eu le grand plaisir et la lourde tâche de m'occuper d'un atelier sur les statistiques.

M Jost, qui avait eu l'occasion de remarquer mon intérêt pour la partie statistique des nouveaux programmes de seconde, m'avait demandé de préparer quelque chose sur la simulation.

Après une première partie introductive qui précisait dans quel esprit il me semblait falloir lire les contenus de ces programmes et quels pouvaient en être les enjeux, je me suis surtout attaché à montrer un certain nombre de situations qui me paraissaient traduire la nouvelle démarche qui nous est proposée.

Il ne s'agit pas d'exemples simples, directement applicables en classe, mais plutôt de pistes à explorer.

En lisant « Prédire n'est pas expliquer » de René THOM, j'étais tombé sur l'idée que la méthode scientifique consistait à remplacer le réel par un virtuel contrôlé. Le titre de l'atelier m'est alors venu assez naturellement : la simulation est l'une des interfaces entre ce réel et ce virtuel contrôlé.

Un des aspects majeurs de la simulation réside en ce qu'elle suppose l'antériorité du modèle sur le réel étudié. Simuler, c'est toujours au départ disposer d'un modèle. La simulation permettra alors de juger de son adéquation au réel étudié. Elle permettra aussi de mettre en évidence des régularités (comme la fluctuation d'échantillonnage) rendues manifestes par le grand nombre d'essais possibles.

Cet article et ceux qui suivront sont issus du document que j'avais préparé pour le 9 juin, mais des modifications ont été apportées, des points retravaillés, des aspects complétés . . .

Bon voyage en statistique.

Comme bien d'autres professeurs, je n'ai jamais eu une très bonne opinion des statistiques enseignées au lycée. Elles font souvent partie, avec la géométrie dans l'espace, des laisser pour compte de la classe de seconde.

Sans épaisseur mathématique apparente, utilisées fréquemment comme un moyen de mettre en confiance les élèves (surtout dans des classes dites « faibles »), ou comme un dispositif permettant la révision des pourcentages et autres nombres rationnels, elles sont oubliées ou détournées ; en tout cas, rarement enseignées pour elles-mêmes.

Bouleversant dans les chaumières mathématiques : voici des programmes qui prévoient des statistiques déroutantes, incontournables (elles doivent représenter un huitième de l'année de seconde et ont toute chance de faire une entrée remarquée en première et terminale S), truffées de mots et de concepts nouveaux : boîte à moustaches, simulation, modélisation, fluctuation d'échantillonnage, fourchette, intervalle de confiance, estimation, loi des grands nombres, sondages . . .

Notons que dans un même temps disparaissent en seconde des notions comme fonction de répartition ou écart-type. La caractéristique de dispersion utilisée sera l'étendue. Plus simple, plus directement accessible, sans aucune virtuosité calculatoire, elle donne une première approche concrète d'une série statistique. Elle permet en outre de porter une plus grande attention aux valeurs extrêmes, et il est certain que, dans la période de catastrophes écologiques ou climatiques que nous venons de traverser, la démarche a gagné en opportunité.

Plus que de grands discours, ce remplacement traduit bien les objectifs poursuivis : il s'agit de donner une perception moins figée des statistiques. Ne nous embarrassons pas d'outils difficiles à justifier et à calculer. Conservons la moyenne et la fréquence qui sont les concepts intuitifs, basiques. Leur simplicité plaide pour elles. Mais évitons d'en faire de simples prétextes à des exercices de calcul numérique.

Car ce qui est visé ici, c'est clairement de poser les fondements de ce que Daniel SCHWARZ appelle dans son remarquable ouvrage *La statistique et le vivant*, une culture de l'incertain.

La statistique (et non pas les statistiques pour reprendre la différence que fait Daniel SCHWARZ), c'est une forme de pensée fondée sur l'idée du risque consenti, c'est aussi sans doute une école de modestie et de patience, modestie devant son impossibilité inhérente à prouver une fois pour toutes une affirmation (« le tabac donne le cancer »), mais en même temps patience dans la constitution du « dossier » qui permettra de prendre des décisions futures (« fumer augmente la probabilité d'avoir un cancer, il vaut donc mieux ne pas fumer »), patience et ténacité car les embûches sont nombreuses pour rendre acceptable notre « intime conviction ».

Nouvelle approche de la causalité, assez éloignée de celle qui a cours dans l'univers de l'algèbre et l'analyse. Comment donc s'y prendre pour donner du sens à cette démarche statistique : le GTD le précise dans cette phrase extraite du document d'accompagnement de janvier 2000 : « l'esprit statistique naît lorsqu'on prend conscience de la *fluctuation d'échantillonnage* ».

Commençons donc par quelques rappels.

Si l'on mesure la fréquence d'une certaine caractéristique sur divers échantillons d'une population bien identifiée, il y a fort à parier que l'on trouvera bon nombre de résultats différents : c'est la *fluctuation d'échantillonnage*. Pourtant à y regarder de plus près, cette variabilité a des lois qui font que le hasard n'est pas simple chaos : si dans une population on considère des échantillons de même taille, les fréquences observées sur ces échantillons se répartissent autour de la fréquence « vraie » (celle de la population) avec une certaine régularité, ce qui permet d'aboutir à la notion de *fourchette*.

Inversant le problème, les statisticiens se servent de cette régularité, qui est calculable, pour estimer à partir d'un échantillon, avec un risque d'erreur fixé, un intervalle (appelé *intervalle de confiance*) qui contient la fréquence « vraie » de la caractéristique, quand elle est inconnue : on aboutit ainsi à l'*estimation* et un peu plus loin aux *sondages*.

Il faut noter que cette « fréquence vraie » n'a souvent aucun sens : si l'on cherche à connaître le pourcentage de français qui aiment le coca-cola, un tel nombre n'existe pas à proprement parler, puisqu'il se modifie à chaque instant (du fait des morts et des naissances par exemple), aucune enquête exhaustive ne pouvant le déterminer pour cette raison même et ne pouvant mettre en évidence sa « lente » évolution au cours des générations.

L'un des objectifs des nouveaux programmes est sans doute de montrer cette *variabilité*, et d'en dégager empiriquement l'aspect « prévisible ». Sans bien sûr qu'il soit nécessaire (et donc obligatoire) d'aborder la notion d'intervalle de confiance.

La démarche expérimentale n'est généralement pas au cœur de nos pratiques mathématiques ; pourtant en l'absence de tout formalisme, seule la répétition d'un grand nombre d'expériences permet de donner une légitimité à la méthode et aux résultats.

Toutefois, il est vain d'espérer répéter cette rencontre de l'humain avec le hasard, des siècles de tâtonnements et d'observations, en quelques manipulations, même bien préparées. Pour qu'un ordre apparaisse dans le désordre apparent de la distribution des fréquences observées, il vaut mieux que les échantillons sur lesquels ces fréquences sont étudiées soient de taille suffisante et en nombre suffisamment important. Comment alors rendre réalisables ces deux contraintes dans le cadre d'un programme et d'une classe ?

La réponse choisie par le GTD est tout à fait naturelle dans le monde des statistiques : elle s'appelle *simulation*. L'ordinateur ou plus modestement la touche « *renom* » d'une calculatrice remplaceront l'expérience vécue du hasard, fournissant rapidement de nombreux résultats.

Mais la simulation ne se réduit pas à être un moyen pratique d'obtenir des résultats en grand nombre. Elle est au cœur de la démarche statistique avec son indissociable compagnon, la *modélisation*. Car, bien sûr, simuler c'est au départ se donner un modèle probabiliste. Substituer au jet du dé le tirage d'un nombre aléatoire sur un ordinateur, c'est par exemple supposer une situation d'équiprobabilité que, certainement, on ne rencontre que tout à fait exceptionnellement avec un dé réel.

Pour paraphraser René THOM, on peut dire que *simuler c'est remplacer le réel par un virtuel contrôlé*. La modélisation et la simulation sont bien sûr déjà totalement à l'œuvre dans les boules et les urnes, ou dans les roulettes chères à Arthur ENGEL. L'ordinateur permet ici d'amplifier la reproduction de l'expérience. Face à un réel mal connu ou trop complexe, le virtuel de l'ordinateur ou de l'urne, situations maîtrisées, permet d'anticiper, d'analyser des résultats reproductibles.

La simulation présuppose donc qu'une *loi de probabilité* soit posée d'emblée. Les raisons qui conduisent à choisir tel ou tel modèle probabiliste relèvent de considérations diverses comme par exemple la symétrie du dé ou de la pièce. Programmer une machine afin de simuler le jet d'une pièce et constater que la fréquence des « faces » sur un grand nombre de lancers virtuels « tend » vers 0,5, puis en déduire que la probabilité d'obtenir « face » est de 1/2 relève donc d'une incompréhension complète de la question. Ce qui est posée c'est la probabilité, rien d'étonnant (du fait de la *loi des grands nombres*) à retrouver une fréquence tendant vers cette probabilité.

Une telle approche qui se veut « *fréquentiste* » conduit ici à prendre l'effet pour la cause.

La simulation met en évidence la relativité de tout modèle. On ne peut pas considérer un modèle comme unique ni comme définitif. Il s'agit donc de se donner des moyens d'en contrôler la validité, tout en sachant que quel que soit le test utilisé, du fait de la fluctuation d'échantillonnage, il faudra accepter ou rejeter le modèle avec un risque d'erreur consenti.

Mathématiquement, c'est à travers le concept de *variable aléatoire* que se constitue la charnière essentielle entre le réel et le modèle, puis entre le modèle et la simulation, pour enfin permettre d'étudier l'écart entre le virtuel simulé et le réel observé ...

Signalons que tout ce qui vient d'être dit est résumé dans le document d'accompagnement : « formellement, simuler une expérience, c'est choisir un modèle pour cette expérience ». Apprendre à simuler, c'est donc apprendre à bâtir un *protocole expérimental* qui permettra de construire une simulation « convenable » de la loi de probabilité que l'on suppose pour l'expérience aléatoire. Ce que l'on attend en seconde c'est semble-t-il que l'élève apprenne à échafauder de tels protocoles sur des cas simples, puis à interpréter les résultats obtenus. Il est clair que l'évaluation d'un tel comportement peut difficilement s'intégrer aux habituels contrôles. C'est pourquoi le GTD préconise un cahier de statistiques, qui accompagnerait l'élève tout au long de l'année (et au delà ...), permettant un suivi individualisé du travail accompli (voir sur ce point le deuxième paragraphe de la page 10 du document d'accompagnement).

À suivre ...