Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

La génération de scénarios aléatoires

par Bernard Beauzamy

Résumé opérationnel

A partir de données recueillies sur un process (ce qu'on appelle un "historique"), nous montrons comment générer des "scénarios" : données prospectives, hypothétiques, mais conformes à la loi de probabilité définie à partir de l'historique. Ceci se fait très simplement, à partir de "macros" en VBA sous Excel ; l'approche présentée ici est beaucoup plus simple que ce préconisent les ouvrages académiques présentant la génération de nombres aléatoires selon une loi donnée.

On peut aussi considérer le cas de lois non-stationnaires, c'est-à-dire dépendant du temps. C'est typiquement le cas du mouvement du conducteur d'une automobile, qui se poserait la question suivante : voici la liste des distances parcourues pendant la première heure ; où serons-nous au bout de la deuxième heure ? Il faut alors travailler sur les vitesses, et non sur les positions.

I. Génération de nombres aléatoires

Si on se réfère aux lois usuelles (loi uniforme, loi de Gauss, etc.), il est très facile de générer des nombres suivant ces lois ; la plupart des logiciels spécialisés sont dotés des fonctions appropriées. Sous Excel, par exemple, il existe une fonction rnd() qui génère des nombres selon la loi uniforme sur l'intervalle $[0,1]$. C'est très commode, mais ne répond en rien aux besoins d'une entreprise, dont la préoccupation s'énonce comme suit : nous disposons d'un historique relatif à certains événements ; nous voudrions simuler un scénario d'événements possibles dans l'avenir.

II. Présentation du besoin

Commençons par bien expliquer ce que recouvre un historique et ce qu'est un scénario.

Nous avons eu l'occasion, récemment, de travailler pour la SNCF : plan d'inspection des rails. A chaque rail inspecté, on note la date de pose, qui se trouve être entre 1850 et 2020. Cela concerne, dans la base de données que nous avons vue, environ 600 000 rails ; à partir de cela on pourra constituer un histogramme : tel pourcentage entre 1850 et 1860, tel entre 1860 et 1870, etc. Cet histogramme est d'un grand intérêt pour la SNCF, car il renseigne sur la proportion de

rails anciens, donc à remplacer en priorité. Il peut être considéré comme une loi de probabilité : si on prend un rail au hasard, voici la probabilité de sa date de pose.

A partir de là, on peut générer un scénario, qui serait du type suivant : nous allons inspecter 100 rails pris au hasard, leurs dates de pose peuvent être 2010, 1870, …., ou ce que l'on voudra. Un tel scénario n'a aucun intérêt pour la SNCF, qui décide des inspections sur critères objectifs, et non au hasard. On voit ici une situation où l'histogramme a un intérêt (la loi de probabilité), mais non les scénarios.

Nous travaillons également pour l'Agence Nationale des Titres Sécurisés ; la question porte sur la délivrance de titres, par exemple la carte d'identité, selon les mois. L'Agence dispose d'un historique et veut disposer d'un scénario : combien de demandes attendues en juin 2024, juillet, etc. Là, au contraire, le besoin en scénarios est réel, parce qu'il faut disposer des ressources (en hommes, en matériel), pour faire face à la demande, qui est très variable.

On comprend bien, ainsi, la différence entre loi de probabilité et scénario. Un constructeur automobile s'intéressera, par exemple, à la proportion d'appels à la garantie sur tel modèle (loi de probabilité), mais pas du tout à un scénario, qui générerait des automobiles particulières (avec leur immatriculation), faisant appel à la garantie.

A l'inverse, une compagnie d'assurances qui assure le risque tempête sera intéressée par un scénario, disant : compte-tenu de l'historique des tempêtes en France métropolitaine, voici un scénario d'apparition des tempêtes sur les dix prochaines années ; un tel scénario lui sera utile, pour voir si elle possède les ressources financières destinées à couvrir les remboursements.

III. Approche mathématique

A. Cas continu

Si la variable X a une densité continue f et si la fonction de répartition F est strictement croissante, alors il suffit de tirer selon une loi uniforme des nombres u entre 0 et 1 et de considérer la suite des $F^{-1}(u)$. En effet, si U désigne une variable suivant une loi uniforme entre 0 et 1, on a :

$$P\left\{F^{-1}(U) \le x\right\} = P\left\{U \le F(x)\right\} = F(x)$$

Dans le document :
https://www.editions.polytechnique.fr/files/pdf/EXT_1616_6.pdf
on présente ce résultat comme dû à Von Neumann ; c'est très peu probable ; comme il est très élémentaire, il était certainement connu de Laplace.

De manière générale, l'ouvrage de référence pour la génération de nombres aléatoires selon diverses lois est le livre :

Non-Uniform Random Variate Generation, Luc Devroye, Springer, 1986.

L'approche valable pour une densité continue ne nous aide pas en pratique : nous avons une suite d'observations, ce qui ne correspond pas à une densité continue, mais à un histogramme. On peut toujours, assurément, passer de l'histogramme à une représentation continue, mais ceci est artificiel et requiert un important travail de préparation.

B. Cas réel

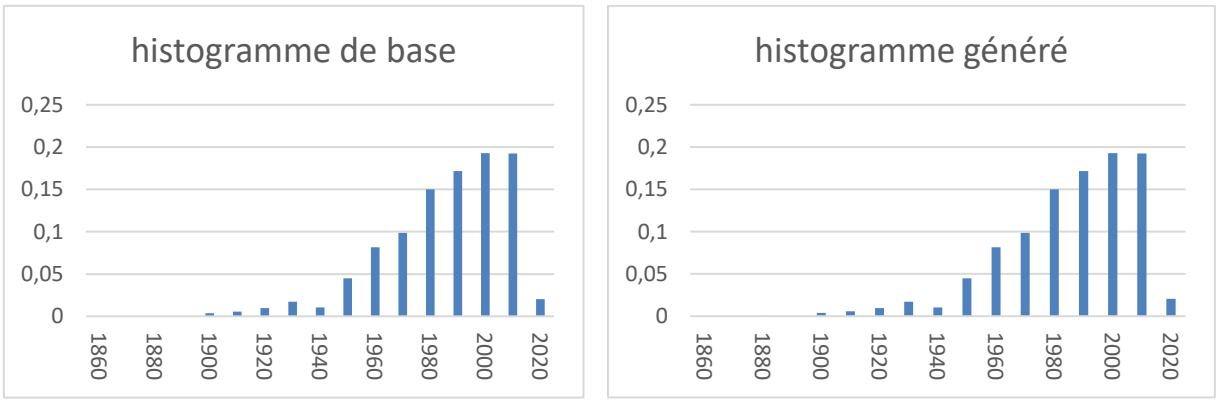
En pratique, nous disposons d'une colonne de N nombres (600 000, pour la SNCF : âge des rails inspectés) ; nous réalisons un histogramme : dans chacun des intervalles $[1850-1860[$, $[1860-1870[$, etc., on met le pourcentage d'observations correspondant. Cet histogramme représente une loi de probabilité et nous voulons générer des nombres suivant cette loi.

Comme nous allons le voir, c'est très facile ; beaucoup plus facile que la méthode donnée pour le cas continu.

Nous commençons par trier les N nombres par ordre croissant (Excel fait cela vite et bien). Ensuite, nous tirons un nombre u selon une loi uniforme entre 0 et 1 (Excel fait également cela vite et bien). Dans la liste triée, le nombre retenu sera celui de numéro $\text{int}(u \times N)$, où $\text{int}(\ )$ désigne la partie entière.

La démonstration est évidente : prenons par exemple le premier intervalle d'âges, $[1850-1860[$. Soit n_1 le nombre d'éléments dans cet intervalle ; la probabilité p_1 est $p_1 = \frac{n_1}{N}$. Par conséquent, si on tire selon une loi uniforme, on tombera dans cet intervalle avec la probabilité p_1, et de même pour les autres intervalles.

En procédant ainsi, on peut générer un nombre quelconque de rails ; mettons 700 000, avec des âges suivant la loi de probabilité donnée par l'histogramme historique. On peut construire l'histogramme à partir de cet échantillon de 700 000 et il sera identique à l'histogramme historique.



(réalisé sans aucun trucage)

Voici les dix dates de pose des rails, générées par ce procédé :

1979, 1980, 1997, 1987, 2014, 1999, 1991, 2001, 2003, 1953, 2008.

On nous accordera que cette information ne présente aucun intérêt.

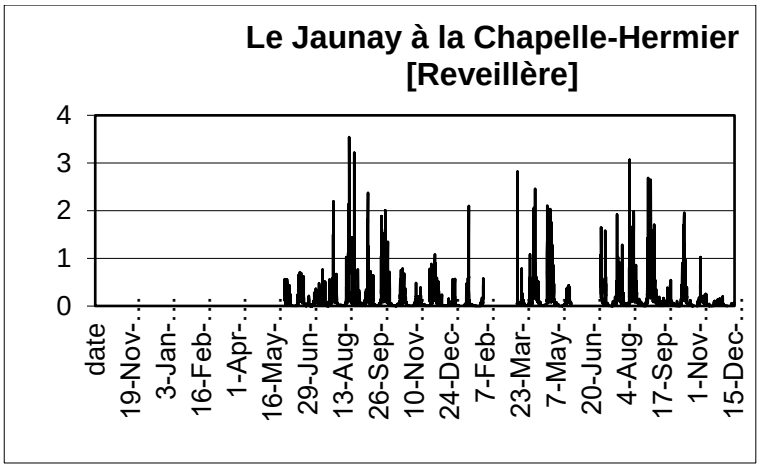
IV. Comparaison méthodologique

Malgré les apparences, l'approche discrète et l'approche continue relèvent de la même idée. Appelons X la variable "date de pose du rail" ; soit f sa densité et F sa fonction de répartition. L'ensemble $\{x, F(x) < p_1\}$, avec $p_1 = \frac{n_1}{N}$ est l'ensemble des rails dont l'âge est dans l'intervalle $[1850-1860[$. Donc cet ensemble est bien $F^{-1}(p_1)$, et ainsi de suite pour les autres intervalles.

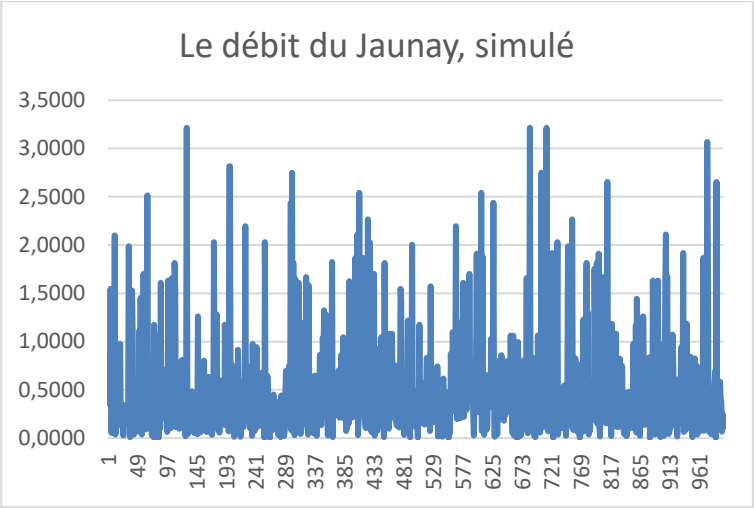
Mais la construction discrète est infiniment plus simple, à la fois conceptuellement et techniquement : il suffit de trier un tableau de données, et on n'a pas besoin de comprendre ce qu'est une fonction de répartition, ni comment on peut l'inverser.

Autres applications

On peut se servir de ce procédé pour simuler l'évolution d'un phénomène naturel. Voici le débit du fleuve Le Jaunay (Bretagne), extrait de notre livre [RDM] :



En se servant des débits relevés, voici une simulation de débits fictifs, pendant une période de 1000 jours :



L'aspect général est bien le même, mais la saisonnalité n'est pas respectée. Elle pourrait l'être, si nous prenions des précautions supplémentaires : générer séparément selon les mois ou les saisons.

V. Evolution temporelle

La génération présentée ci-dessus n'est valable que si la loi de probabilité, dans l'avenir, est la même que dans le passé ; on dit alors que la loi est "stationnaire". Ce n'est pas le cas, en particulier, pour les phénomènes qui impliquent un mouvement, si on se contente de regarder les mesures de position.

Illustrons ceci par un exemple simple : soit une automobile, dont on enregistre les positions toutes les 5 minutes, pendant une heure. Ces positions sont décrites par des points sur un axe, entre 0 et 100 (graduations en km, en admettant que l'automobile ne dépasse pas 100 km/h). Si nous utilisons ces données de position pour générer les valeurs suivantes (l'heure qui suit), elles seront évidemment entre 0 et 100 sur le même axe : on a l'impression que l'automobile n'avance pas.

Il faut en réalité travailler sur les vitesses, de la manière suivante :

Supposons que la vitesse, par tranche de 5 minutes, soit donnée par le tableau suivant :

vitesse	position	temps
98	0,00	0
24	8,17	5
53	10,17	10
10	14,58	15
99	15,42	20
67	23,67	25
1	29,25	30
57	29,33	35
10	34,08	40
10	34,92	45
79	35,75	50
28	42,33	55
	44,67	60

On en déduit la position pour chaque tranche de 5 minutes ; au bout d'une heure, l'automobile aura parcouru 44,67 km.

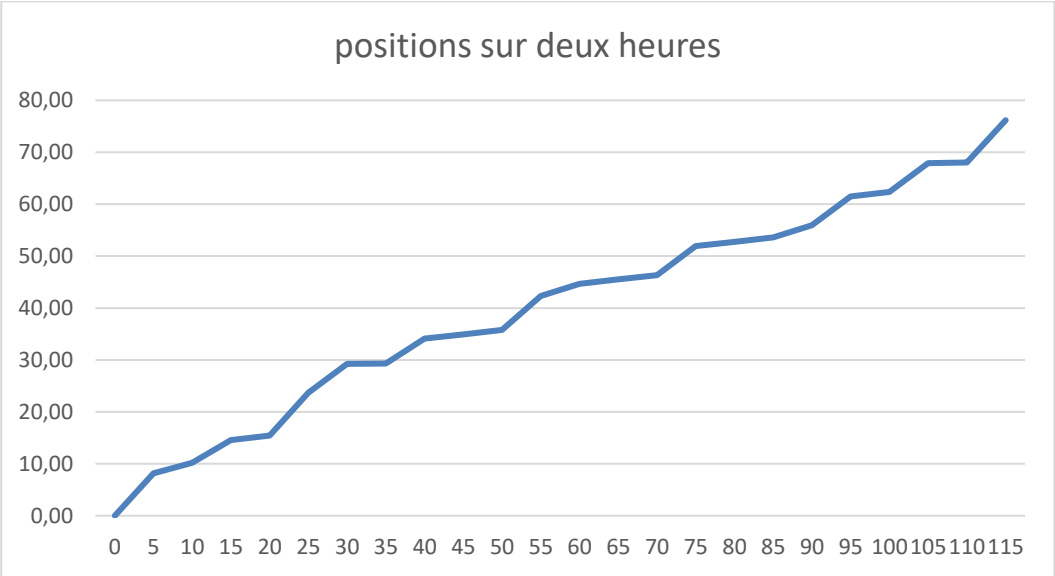
On génère ensuite 12 vitesses, selon le principe ci-dessus, et on obtient :

vitesse générée
79
10
10
67
10
10
28
67
10
67
1
98

On en déduit la position, par tranche de 5 minutes, pour l'heure suivante ; voici le résultat obtenu pour deux heures :

position	temps
0,00	0
8,17	5
10,17	10
14,58	15
15,42	20
23,67	25
29,25	30
29,33	35
34,08	40
34,92	45
35,75	50
42,33	55
44,67	60
45,50	65
46,33	70
51,92	75
52,75	80
53,58	85
55,92	90
61,50	95
62,33	100
67,92	105
68,00	110
76,17	115

et sous forme de graphique :



Le raisonnement fait mérite d'être explicité en français usuel. Il consiste à dire que, pendant la seconde heure, les vitesses rencontrées suivront la même loi de probabilité que pendant la première heure, peut-être pas dans le même ordre. Par exemple, la vitesse 10 km/h, rencontrée

trois fois pendant la première heure, a une probabilité $\frac{3}{12}=\frac{1}{4}$, et de même pour les autres valeurs.

Rappel : comment construire un histogramme ?

L'utilisateur choisit la valeur a, borne inférieure de la plus petite classe et w (width), taille de chaque classe. Les classes sont alors de la forme $\left[a+kw, a+\left(k+1\right)w\right[$, $k=0,1,...$; la borne de gauche est incluse, la borne droite exclue, par convention.

Un nombre x appartient à la $k^{ème}$ classe si :

$$a+kw \le x < a+\left(k+1\right)w$$

ou encore :

$$k \le \frac{x-a}{w} < k+1$$

Par conséquent, le numéro de la classe où se trouve x est $k=\text{int}\left(\frac{x-a}{w}\right)$, où int() désigne la partie entière.