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

Durée de retour

et probabilité d'occurrence

par Bernard Beauzamy

Juillet 2022

Initialement rédigé en complément méthodologique à un contrat avec l'Agence Nationale des Déchets Radioactifs (Andra), 2022.

I. Durée de retour d'un phénomène

Il s'agit d'un concept très simple, qui n'a rien de probabiliste en soi. On désigne ainsi le temps moyen qui s'écoule entre deux occurrences d'un phénomène. Si celui-ci s'est produit aux instants $t_1,...,t_{N,}$ la durée de retour est :

$$dr = \frac{1}{N-1}\left(\left(t_2 - t_1\right) + \left(t_3 - t_2\right) + \cdots + \left(t_N - t_{N-1}\right)\right) = \frac{t_N - t_1}{N-1}. \quad (1)$$

Le mot "durée" est impropre : il vaudrait mieux dire "intervalle de temps moyen", mais il est consacré par l'usage. Le concept n'a rien de probabiliste : la durée de retour de la Comète de Halley est de 76 ans. On peut limiter la durée d'observation si on le souhaite, et dire par exemple : la durée de retour de tel phénomène entre l'an 1000 et l'an 1500 est de tant.

A l'évidence, la condition nécessaire et suffisante pour que la durée de retour puisse être définie est que le phénomène se soit produit au moins deux fois. D'où la règle :

Règle : on s'interdira de parler de durée de retour pour un phénomène s'il ne s'est jamais produit, ou bien s'il s'est produit une seule fois.

II. Loi de probabilité

Il s'agit là d'un concept beaucoup moins simple qu'on ne l'imagine habituellement. Considérons un magasin de vêtements, et notons l'âge des visiteurs, par classe de 10 ans (0-10, 10-20, …, 90-100). Nous ne savons pas quel sera l'âge du prochain visiteur ; Dieu le sait mais ne nous le dira pas. Nous pouvons nous en faire une idée si le magasin fonctionne depuis suffisamment longtemps et si nous avons consciencieusement enregistré les âges. Nous notons p_1 la proportion de visiteurs de la première tranche (0-10) : $p_1 = \frac{n_1}{N}$, où n_1 est le nombre total de visiteurs ayant leur âge dans la première tranche et N le nombre total de visiteurs, et ainsi de suite pour les autres tranches. Si à chaque tranche on associe son âge moyen (entre 10 et 20 : 15), nous avons 10 valeurs, notées x_i, $i = 1,...,10$ (à savoir 5, 15,…, 95) et 10 probabilités, notées p_i, avec $p_i \geq 0, \quad \sum_{i=1}^{10} p_i = 1.$ A ce stade, l'information n_i, N s'est perdue.

Les valeurs que peut prendre l'âge (à savoir 5, …, 95) sont les valeurs que peut prendre une "variable aléatoire" et les p_i constituent la loi de probabilité de cette v.a. Cette information ne remplace pas celle que Dieu refuse de nous fournir (âge du prochain visiteur), mais c'est tout ce dont nous disposons et il va falloir s'en contenter. Elle est tout de même porteuse d'enseignement, en particulier elle permet de savoir quelles sont les tranches d'âge les plus représentées et de disposer des stocks en conséquence. En langage standard, on peut regarder p_i comme la proportion de la $i^{ème}$ tranche d'âge au sein de la population constituée de tous les visiteurs.

La loi de probabilité ne sait pas faire de différence entre un magasin récent et un magasin ancien, puisque seuls les quotients $\frac{n_i}{N}$ sont conservés. En pratique, ceci est source de difficultés : les enseignements portés par un magasin récent sont beaucoup moins fiables, en particulier si le magasin vient juste d'ouvrir.

La loi de probabilité est dite "stationnaire" si les p_i restent toujours les mêmes au cours du temps. La clientèle ne vieillit pas, ne rajeunit pas. Les exemples pratiques de lois stationnaires sont très rares : dans la pratique, les répartitions évoluent. Un exemple évident de loi stationnaire est donné par le jeu de Pile ou Face : la probabilité de chaque résultat est $1/2$ quelles que soient les circonstances. Pour le reste, les lois ne sont pas stationnaires : les populations vieillissent ou rajeunissent, les appareils s'usent, etc.

Prenons un exemple concret, quoiqu'académique, lié à la température d'un système. Un système physique (où l'on mesure la température) aura une loi stationnaire s'il enferme une grosse sphère, à l'intérieur de laquelle se trouvent de petits bulletins. Sur chaque bulletin, on a inscrit une température. Chaque jour, un robot ouvre la sphère, prend un bulletin au hasard, règle la température selon l'indication du bulletin et remet le bulletin dans la sphère. Cette loi est stationnaire, parce que la sphère est constamment la même, de même que le protocole de lecture. Naturellement, la température n'est pas constante. On note p la probabilité d'une température donnée, un jour donné.

Notons K le nombre de températures différentes. Supposons qu'un jour donné, la température 34°C soit apparue. Quelle est la probabilité de la revoir au bout de k jours, mais non avant ? C'est $(1-p)^{k-1}\,p$. Quelle est l'espérance de ce temps d'attente ? C'est $E = p\sum_{k=1}^{+\infty} k(1-p)^{k-1} = 1/p.$

On a démontré un théorème : la durée de retour (durée moyenne du temps d'attente avant de revoir une étiquette) est égale à l'inverse de la probabilité de la température considérée.

III. Deux concepts différents : source de confusion

Dans le cas de notre magasin de vêtements, nous mettons en évidence deux concepts de probabilité, absolument différents :

- Une loi de probabilité, qui caractérise la proportion de chaque tranche d'âge parmi tous les visiteurs. On dira par exemple que les 0-10 ans (représentés par leur valeur médiane 5 ans) représentent 6% des visiteurs. Nous avons donc un concept clair et bien défini : une liste d'âges et une liste de proportions. Ces proportions ont vocation à être fixes, si le système est stationnaire, ce que l'on peut supposer (du moins sur une période de quelques années). Nous l'appellerons "probabilité descriptive".

- On peut parler de durée de retour pour chaque tranche d'âge : ce sera l'intervalle de temps moyen (il faut préciser l'unité) séparant la visite de deux personnes de même âge. Si les âges des visiteurs sont enregistrés tous les jours, l'unité sera la journée. Dans le cas du magasin pris pour exemple, pour les 5 ans, nous avons 16 visites sur un total de 82 jours : la durée de retour moyenne est donc de 5.125 jours (en langage imagé, on voit un bambin de 5 ans tous les 5 jours, ou un peu plus). La probabilité, par jour, de voir un bambin de 5 ans est donc $p = \frac{1}{5.125} \approx 0.195.$ Nous l'appellerons "probabilité d'occurrence".

Cette seconde probabilité :

- N'a rien à voir avec la première ;
- Suppose que le système soit stationnaire ;
- Fait nécessairement référence à un intervalle de temps (ici la journée).

Elle est d'usage fréquent dans l'industrie ; on peut l'appeler "probabilité d'occurrence d'un événement redouté". Un industriel se demandera par exemple : quelle est la probabilité que mon usine tombe en panne au cours de l'année qui vient (période de 12 mois) ? L'intervalle de temps moyen séparant deux pannes, c'est-à-dire la durée de retour de l'événement "panne", selon la terminologie du premier paragraphe, s'appelle "Mean Time Between Failure" (en abrégé MTBF) dans la terminologie des experts en fiabilité.

La probabilité d'occurrence d'un événement redouté se ramène au cadre classique des probabilités par l'introduction (factice) de la loi des grands nombres. On imaginera 1 000 usines, toutes du même modèle, fonctionnant indépendamment, et on se demandera : combien auront manifesté une panne dans les 12 mois à venir ? Si cette introduction de la loi des grands nombres ne peut être faite ou n'a pas de sens, c'est que le problème ne peut être traité de manière probabiliste.

IV. Très faible probabilité

Fixons-nous un seuil, mettons 10^{-6} par an, et posons-nous la question : pouvons-nous parler d'événements climatiques, à Bure (petite ville de l'est de la France, où l'Andra prévoit d'installer un site de stockage souterrain) ayant cette probabilité ? Tout d'abord, il est évident qu'il n'y a qu'un seul Bure, et on ne peut pas multiplier les Bure comme on multiplie les usines. En revanche, on peut étendre (de manière factice) l'intervalle d'étude : convenir que notre surveillance dure un milliard d'années. Dans ces conditions, la question a un sens : un événement de probabilité 10^{-6} par an se rencontrera environ 1 000 fois en un milliard d'années.

En pratique, ce qu'on enregistre est une température moyenne par jour (on conserve le maxima de ces températures). Si l'événement est "dépasser 41°C", ou ce que l'on voudra, il faut que cette température ait été dépassée pendant toute une journée : une durée de 20 minutes ne suffira pas, et la station de mesure n'indiquera rien. Si on travaille sur une année seulement, la probabilité la plus faible que l'on peut détecter est $\frac{1}{365}$, si on travaille sur 20 ans, la probabilité la plus faible est $\frac{1}{365^{20}}$ etc.

Peut-on apporter une réponse mathématique, même grossière, à la question : quelle est la probabilité que la température atteinte à Bure, pendant l'année 2025, soit 45,3°C ? Il s'agit d'un futur proche et d'une valeur limite proche de celle qui a déjà été rencontrée. La réponse est que ceci est impossible en général, même si on dispose des enregistrements sur un grand nombre d'années et même si le futur est très proche. Voici néanmoins deux réponses, toutes deux très imparfaites.

Une première possibilité est d'utiliser :

A. La formule de Laplace

Elle s'écrit sous la forme (voir le livre [MPPR]) :

$$P(n_1) = \frac{N+1}{N+N_1+1} \frac{\binom{N_1}{n_1}\binom{N}{n}}{\binom{N+N_1}{n+n_1}}. \qquad (2)$$

Sachant qu'on a enregistré n accidents sur N essais, la formule donne la probabilité d'enregistrer, dans le futur, n_1 accidents sur N_1 essais. Ici, on a fait $N =$ 7 463 mesures de la température (depuis 2001/06/21, il y a des trous) et on a $n =$ 0 accidents, en comptant comme "accident" le fait de dépasser la température limite de 40.6°C. La probabilité de ne pas dépasser la température limite en $N_1 =$ 1 000 jours d'observations sera donc :

$$p(0) \approx 0.88$$

Le résultat serait le même pour toute température limite, par exemple si l'on fixait 45°C, ce qui est peu satisfaisant.

B. L'Hypersurface Probabiliste

On peut aussi utiliser la théorie dite "EPH" (Expérimental Probabilistic Hypersurface), due à Olga Zeydina et Bernard Beauzamy (voir le livre [PIT]). Etant donné un ensemble d'observations (les relevés de température dans le passé), l'EPH les "propage" sous forme d'une loi de probabilité en tout point de l'avenir ; l'EPH dira par exemple : voici la densité de probabilité, pour la température, attendue pour le 23 février 2025. Mais l'EPH travaille sur un concept totalement abstrait d'information, et n'incorpore aucune spécificité qui peut être liée au fait qu'il s'agit ici de températures.

V. Du mauvais usage des probabilités

La langue de tous les jours fait un usage abusif des probabilités "il est probable que", mais il faut se montrer tolérant. Dans les articles scientifiques, en revanche, on s'efforcera de vérifier que le cadre est bien celui que les définitions autorisent.

En l'absence de définition claire de la loi de probabilité sur laquelle on souhaite travailler, les éléments de langage les plus simples deviennent source de confusion et de querelles. Il est donc préférable de bien clarifier les données, les hypothèses et les concepts.

Pour beaucoup de gens, cependant, l'utilisation d'un vocabulaire probabiliste inapproprié est une façon de masquer la faiblesse des raisonnements. Ceci n'est pas nouveau, et le meilleur exemple en est donné par l'Affaire Dreyfus (1894), où Bertillon a monté de toute pièce un "système probabiliste" destiné à montrer que Dreyfus avait écrit le bordereau. Lorsque, en 1903, la Chambre Criminelle de la Cour de Cassation a demandé à Henri Poincaré de porter un jugement critique sur le "système Bertillon", celui-ci a rédigé un rapport cinglant, où il écrit notamment "le calcul des probabilités ne devrait pas empêcher les savants d'avoir du bon sens" [Dreyfus].

Même si l'erreur Dreyfus est de loin la pire de toutes, des erreurs semblables sont encore commises de nos jours. En 1999, en Angleterre, une femme a été condamnée à la prison à vie, sur la base d'une erreur commise par un statisticien dans un raisonnement de nature probabiliste [https://fr.wikipedia.org/wiki/Sally_Clark]. A l'époque, l'Institut de Recherche Criminelle de la Gendarmerie Nationale avait consulté la SCM : que faire pour que ceci ne se produise pas en France ? Notre recommandation n'a pas varié : le raisonnement probabiliste doit, par sa nature même, être banni des affaires criminelles.

Références

[MPPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. SCM SA, ISBN 2-9521458-0-6. ISSN 1767-1175, broché, 369 pages. Mars 2004. Seconde Edition, juin 2016.

[RDM] Bernard Beauzamy, Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes. SCM SA, ISBN 2-9521458-2-2, ISSN 1767-1175.

[PIT] Olga Zeydina - Bernard Beauzamy : Probabilistic Information Transfer. SCM SA, ISBN 978-2-9521458-6-2, ISSN 1767-1175, relié, 208 pages. Avril 2013.

[Dreyfus] http://images.math.cnrs.fr/Des-mathematiciens-dans-l-affaire-Dreyfus.html

Affaire Sally Clark : https://fr.wikipedia.org/wiki/Sally_Clark