Prévoir les ressources nécessaires pour atteindre son but :

Le cas d'espèce de la sonde Rosetta

Dans les heures et les jours qui ont suivi son atterrissage historique sur la comète 67P/TG, le 12 novembre 2014, le module Philæ, largué par la sonde européenne Rosetta, a transmis de nombreuses images et données de forage qui permettront à terme de mieux comprendre la genèse du système solaire.

Christian Genest
James A. Hanley
Université McGill

Cet exploit scientifique, réalisé à 500 millions de kilomètres de la Terre, est l'aboutissement d'un périlleux voyage entrepris par la sonde Rosetta le 2 mars 2004. La mission, qui a nécessité 10 ans de préparatifs et des investissements évalués à 1,3 milliards d'euros, représente aussi un défi technique de grande taille.

Dans son périple de plus de 6,5 milliards de kilomètres, qui a duré 10 ans, la sonde a dû résister aux plans thermique et énergétique aux grandes variations d'amplitude de l'éclairage solaire imposées par sa trajectoire. Ses composants électroniques ont aussi été exposés à de fortes doses de radiations dues aux éruptions solaires.

La question de la fiabilité des équipements et l'optimisation du matériel de secours se pose dans toute mission sans ravitaillement. Elle est d'autant plus cruciale dans les voyages cosmiques, soumis à de fortes contraintes de poids et d'espace. Il faut éviter de dédoubler inutilement les systèmes mais s'ils flanchent, on court à l'échec.

Puisque les pannes sont imprévisibles — et donc de nature stochastique — l'optimisation du nombre de pièces de rechange en cargo s'appuie sur des calculs de probabilités. Nous nous proposons ici d'illustrer la chose en nous inspirant librement de l'ambitieux projet Rosetta de l'Agence spatiale européenne.

Si les ressources m'étaient comptées...

Imaginons qu'une certaine pièce d'équipement électronique essentielle au bon fonctionnement de la sonde Rosetta s'use si lentement qu'en pratique, seul le flux de plasma d'un vent solaire puisse l'affecter. Quand cette pièce tombe en panne, elle est immédiatement relayée par une pièce identique, et ainsi de suite jusqu'à épuisement du stock. Si on dispose au départ de k pièces, le système est donc opérationnel en continu jusqu'à l'occurrence de la k-ième panne.

Concrètement, supposons qu'une rafale de vent solaire secoue la sonde en moyenne une fois tous les 16 mois, à des moments aléatoires (imprévisibles) et indépendants les uns des autres. On s'attendrait donc à avoir en moyenne $\lambda = 3/4$ de panne par an, soit $\mu = \lambda \times 10 = 7{,}5$ pannes en 10 ans. Si la sonde n'avait que $k = 3$ pièces, il y aurait alors peu de chances qu'elle soit encore opérationnelle lors du rendez-vous avec la comète 67P/TG (« Tchouri » pour les intimes).

Plus généralement, quelles seraient les chances de succès de Rosetta si elle disposait au départ de 6, 9, 12 ou 18 pièces ? Pour quantifier le risque de panne en fonction du nombre total de pièces, on doit formuler des postulats, soit sur la distribution du nombre de pannes, soit sur la durée de vie de chacune des pièces. C'est ici que la théorie des probabilités et la statistique volent au secours du décideur.

Deux modèles stochastiques complémentaires

Soit N le nombre inconnu de pannes à survenir au cours d'une mission d'une durée de t ans. Il est raisonnable de supposer que cette variable obéit (à peu près) à une loi dite « de Poisson », du nom du mathématicien français Siméon Denis Poisson (1781–1840) qui l'a proposée. Ce modèle stipule que si $\mu = \lambda t$ est le nombre moyen de pannes au cours d'une mission de t ans, alors la probabilité (Pr) que $N = n$, c'est-à-dire la probabilité que l'on observe n pannes, est donnée en tout $n \in \{0, 1, \ldots\}$ par

$$\Pr(N = n) = \frac{\mu^n e^{-\mu}}{n!},$$

où $n! = 1 \times 2 \times \cdots \times n$ pour $n \in \{1, 2, \ldots\}$ et $0! = 1$ par convention.

Si on dispose au départ de k pièces, une mission de t ans sera complétée avec succès à condition de subir au plus $k - 1$ pannes. La probabilité de cet événement est

$$\Pr(N \leq k-1) = \Pr(N=0) + \Pr(N=1) + \cdots + \Pr(N=k-1).$$

Quand $\lambda = 3/4$, $t = 10$ et $k = 12$, cette probabilité est d'environ 0,921 (ou 92,1%), comme on peut le calculer avec le tableur Excel en utilisant la commande

LOI.POISSON(11;7.5;cumulative=VRAI).

Une autre façon de procéder, équivalente et complémentaire, serait de mesurer la durée de vie totale du système. Si le module est muni de k pièces et qu'on note V_i la durée de vie de la i-ième d'entre elles, la durée de vie totale est alors donnée par

$$V = V_1 + \cdots + V_k$$

et la mission sera un succès si $V > t$ ans. On doit donc forcément avoir

$$\Pr(V > t) = \Pr(N \leq k - 1),$$

ce qui laisse soupçonner un lien entre la loi discrète du nombre N de pannes et la loi continue du temps de survie V du système.

Puisque les pièces sont identiques, il est raisonnable de supposer que leurs durées de vie $V_1, \ldots, V_k$ sont mutuellement indépendantes et de même loi. Il se trouve que si le nombre N de pannes est une variable de Poisson dont la valeur moyenne est $\mu = \lambda t$, on a alors

$$\Pr(V_i > t) = e^{-\lambda t},$$

pour tous $i \in \{1, \ldots, k\}$ et $t > 0$. En effet, sachant que $0! = 1$, on voit que

$$\Pr(V_i > t) = \Pr(V_1 > t) = \Pr(N = 0) = e^{-\mu} = e^{-\lambda t}.$$

On dit que les variables $V_1, \ldots, V_k$ sont exponentielles de taux λ pannes/an. La durée de vie moyenne de chaque pièce est alors $1/\lambda$ et puisqu'elles sont au nombre de k, la durée de vie espérée du système est k/λ. C'est bon à savoir, mais ce qui intéresse surtout les planificateurs de la mission, c'est $\Pr(V > t)$.

Le mathématicien danois Agner Krarup Erlang (1878–1929), qui s'est beaucoup intéressé à la théorie des files d'attente et à la gestion des réseaux téléphoniques, a déterminé la loi de probabilité de la somme $V = V_1 + \cdots + V_k$ de k variables mutuellement indépendantes de lois exponentielles ayant pour taux λ. Il a montré que si l'on pouvait observer un nombre infini de valeurs de V et qu'on en traçait l'histogramme, la forme de la courbe serait alors donnée en tout $v > 0$ par la formule

$$f(v) = \frac{\lambda(\lambda v)^{k-1} e^{-\lambda v}}{(k-1)!}.$$

La probabilité que $V > t$ est alors donnée par l'aire sous cette courbe entre les points t et l'infini, c'est-à-dire

$$\Pr(V > t) = \int_t^{\infty} f(v)\,dv.$$

Démonstration de l'identité

La preuve se fait par récurrence sur le nombre k de pièces. Notons d'abord que l'identité est vraie lorsque $k = 1$ puisque $0! = 1$, de sorte que

$$\int_{\mu}^{\infty} \frac{w^0 e^{-w}}{0!}\,dw = e^{-\mu} = \Pr(N = 0) = \Pr(N \le 0).$$

Supposons ensuite que la relation soit vérifiée pour un entier k quelconque. Pour montrer qu'elle reste vraie en $k + 1$, il suffit d'intégrer par parties. Sachant que $-e^{-w}$ est une primitive de e^{-w} et que $\frac{d}{dw}(w^k) = kw^{k-1}$, on trouve

$$\int_{\mu}^{\infty} \frac{w^k e^{-w}}{k!}\,dw = \frac{\mu^k e^{-\mu}}{k!} + \int_{\mu}^{\infty} \frac{kw^{k-1} e^{-w}}{k!}\,dw.$$

Le premier terme de la somme est $\Pr(N = k)$ et le second est $\Pr(N \le k\text{-}1)$ par l'hypothèse de récurrence. L'intégrale est donc égale à

$$\Pr(N = k) + \Pr(N \le k\text{-}1) = \Pr(N \le k),$$

tel qu'énoncé. Ceci prouve que le résultat est valide pour tout entier $k \in \{1, 2, \ldots\}$.

La loi de Erlang étant un cas spécial de la loi Gamma, le complémentaire de cette intégrale peut être évalué avec la commande d'Excel :

LOI.GAMMA(t;12;4/3;cumulative=VRAI)

Quand on pose $t = 10$, Excel répond 0,079, dont le complément est bien $1 - 0{,}079 = 0{,}921$.

Un lien utile entre le discret et le continu

En effectuant le changement de variables $w = \lambda v$ dans l'intégrale ci-dessus, et en tenant compte du fait que $dw = \lambda dv$, on trouve

$$\Pr(V > t) = \int_{\mu}^{\infty} \frac{w^{k-1} e^{-w}}{(k-1)!}\,dw.$$

Puisque $\Pr(V > t) = \Pr(N \le k - 1)$, on déduit que

$$\sum_{0}^{k-1} \frac{\mu^n e^{-\mu}}{n!} = \int_{\mu}^{\infty} \frac{w^{k-1} e^{-w}}{(k-1)!}\,dw.$$

Cette identité remarquable, valide pour tout entier $k \in \{1, 2, \ldots\}$, peut également être démontrée directement par récurrence (voir encadré ci-contre). Elle établit un lien direct et précis entre une intégrale et une somme finie. On peut donc évaluer l'une ou l'autre indifféremment, selon les moyens de calcul dont on dispose.

Au début du 20e siècle, il n'y avait pas d'ordinateurs et encore moins de tableur Excel. Dans ses travaux précurseurs sur les tests d'ajustement statistique, le mathématicien anglais Karl Pearson (1857–1936) a donc fréquemment dû évaluer de longues sommes de probabilités de Poisson. Avec le temps, des tables des valeurs de l'intégrale de la loi de Erlang ont été bâties en fonction de k et de μ. Un autre père fondateur de la statistique moderne, l'Anglais Sir Ronald Fisher (1890–1962), a souvent exploité l'identité démontrée dans l'encadré pour s'éviter de fastidieux calculs dans ses travaux.

Encore aujourd'hui, Excel et de nombreux autres logiciels de traitement statistique font implicitement appel à cette identité, car les méthodes de calcul numérique actuelles permettent d'évaluer une intégrale avec grande précision, tandis que les erreurs d'arrondi associées à l'évaluation des différents termes d'une somme se cumulent ; si la somme comporte un grand nombre de termes, le cumul des erreurs peut s'avérer non négligeable même si chacune d'entre elles est très petite.

Autres applications et généralisations

Le modèle Poisson-exponentiel présenté ici trouve beaucoup d'applications, non seulement en aéronautique et en théorie des files d'attente, mais dans de très nombreux domaines, tels que l'assurance, la biochimie ou l'étude des maladies.

Les conditions d'application du modèle sont celles du « processus de Poisson. » Bien qu'elles soient relativement peu contraignantes, elles ne sont pas toujours vérifiées en pratique. Cela se produit entre autres lorsque le nombre de zéros observés est anormalement élevé ou quand les durées de vie des pièces successives ne sont ni exponentielles, ni indépendantes les unes des autres. Si une tempête solaire particulièrement violente était susceptible d'endommager toutes les pièces de rechange de Rosetta simultanément, par exemple, les calculs de probabilité présentés ici ne seraient plus valables. Diverses généralisations et variantes du modèle Poisson-exponentiel permettent de tenir compte de telles particularités.

Le gratuiciel R

R est un langage de programmation et un environnement logiciel adapté au calcul statistique et à l'analyse de données. Doté d'excellentes capacités graphiques en 2 et 3 dimensions, ce logiciel est libre de droits et peut donc être téléchargé gratuitement à partir du site suivant:

http://www.r-project.org/

Il est compatible avec les systèmes d'exploitation UNIX, Windows et MacOS.

Le logiciel R est l'œuvre collective de la communauté statistique internationale. Il est en constante évolution, grâce aux centaines de statisticiens et de programmeurs qui contribuent sans cesse et de façon bénévole à l'amélioration de son contenu, notamment par l'ajout de nouveaux modules reflétant les plus récentes percées au plan de l'analyse statistique. Ce logiciel est très largement utilisé, tant aux fins d'enseignement que de recherche.

Les commandes nécessaires à la production de la figure des pages suivantes sont trop complexes pour être reproduites ici (nous vous les fournirons sur demande), mais si vous tapez

```
ppois(11,7.5) ou
pgamma(10,12,scale=4/3,lower.
    tail=FALSE)
```

sur la ligne de commande, le logiciel répondra 0,9207587 dans les deux cas. À la lecture de l'article, pouvez-vous dire pourquoi ?

Pour de plus amples renseignements concernant le gratuiciel R et ses multiples utilisations, voir par exemple l'ouvrage intitulé *Le logiciel R : Maîtriser le langage – Effectuer des analyses statistiques*, publié en 2011 chez Springer France (ISBN 978-2-8178-0114-8). Ce beau livre, finaliste du Prix Roberval, est l'œuvre de trois professeurs de statistique, Pierre Lafaye de Micheaux (Université de Montréal), Rémy Drouilhet (Université Pierre Mendès France, Grenoble) et Benoît Liquet (The University of Queensland, Australie).

Une figure parlante

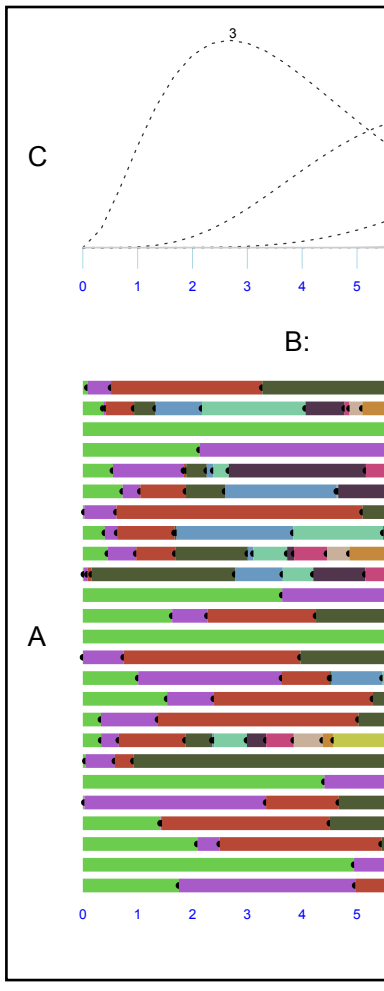
Tous les concepts présentés dans l'article peuvent être illustrés joliment au moyen de la figure ci-contre, produite à l'aide du gratuiciel statistique R [voir encadré n° 2]. La figure est un assemblage de cinq planches étiquetées A, A', B, C et D. Voyons ce qu'elles représentent à tour de rôle :

La planche A comporte 25 lignes composées de segments colorés. Chaque ligne représente une réalisation possible du périple entrepris par la sonde spatiale Rosetta sur une période de 10 ans. Les longueurs des segments colorés représentent les durées de vie successives des pièces, en supposant qu'au départ, la sonde ait à son bord 12 pièces et que les pièces aient des durées de vie mutuellement indépendantes et exponentielles de taux $\lambda = 3/4$ panne/an. Le nombre qui apparaît au bout de la ligne est le numéro de la pièce en fonction au moment du rendez-vous historique entre le module Philæ et la comète Tchouri.

Dans cette simulation réalisée avec R, le hasard a fait que la mission s'est avérée un succès dans 22 cas sur 25, soit 88%. Les scénarios n° 8, 18 et 24 (en comptant à partir du bas) se sont soldés par un échec. La proportion de succès observée est donc assez proche de la valeur prédite par la théorie, soit 92,1%. Pas mal !

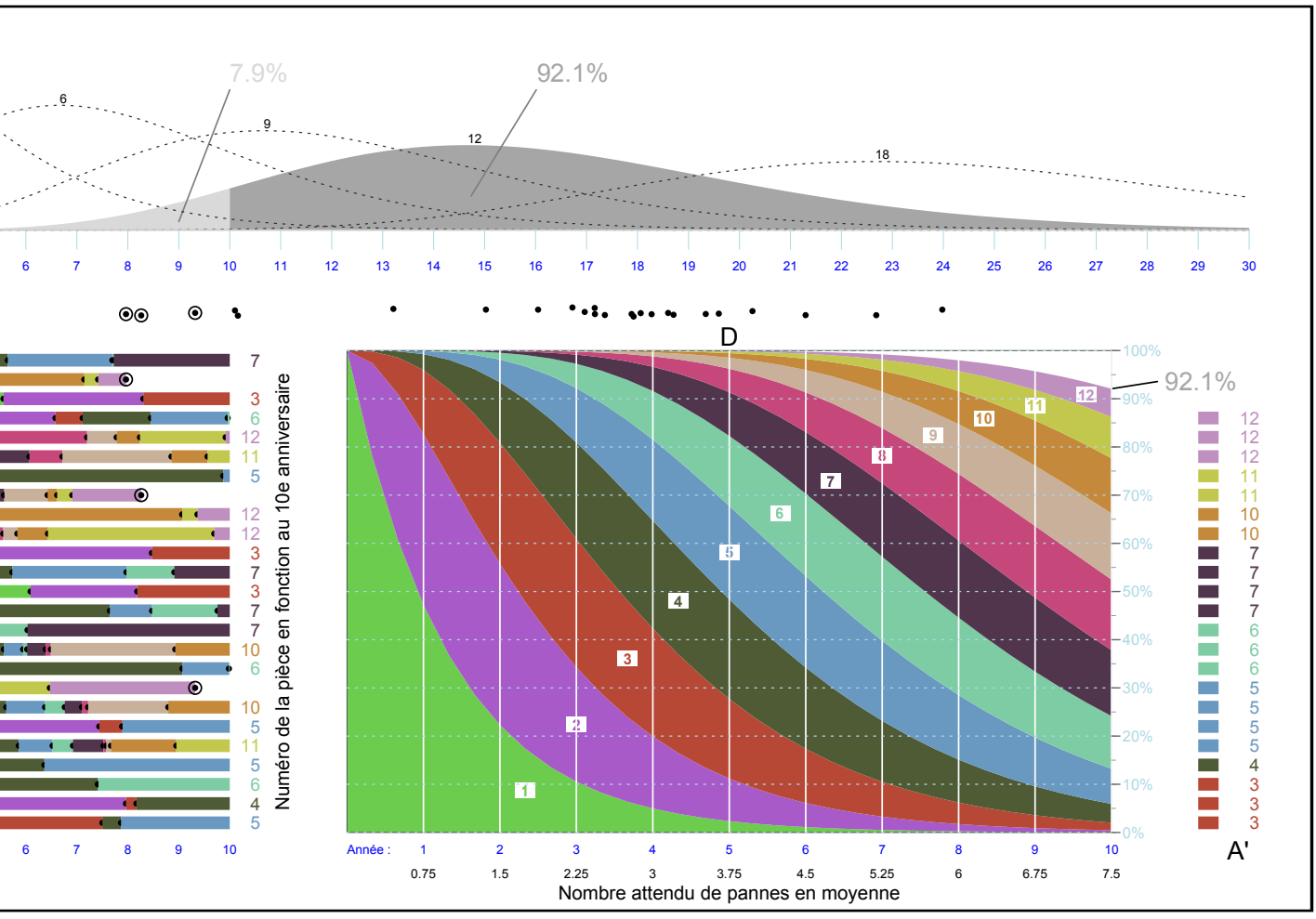
La planche B indique, sur la même échelle de temps que la planche A, le moment où la sonde Rosetta aurait fini par tomber en panne faute de pièces de rechange dans les 25 scénarios simulés. Les trois premiers points sont ceux qui mènent à l'échec des missions simulées n° 8, 18 et 24. Dans le meilleur scénario, la sonde Rosetta ne serait devenue dysfonctionnelle qu'au bout de 24 ans.

Alors que 25 scénarios simulés ont suffi pour obtenir une estimation somme toute acceptable de la probabilité de défaillance pour une mission de 10 ans, la dispersion des points le long de la droite de la planche B donne une idée assez grossière de la distribution de la variable V, la durée de vie totale du système. Il est plus difficile d'estimer une courbe que d'estimer un nombre ! Pour mieux appréhender la loi, il faudrait faire des milliers et des milliers de répétitions et lisser l'histogramme des valeurs obtenues. C'est souvent l'approche retenue par les statisticiens dans les situations très complexes, faisant intervenir de nombreuses variables. Dans le cas présent, il se trouve qu'un calcul exact est possible, comme on l'a déjà expliqué.



L'histogramme idéalisé (la fonction $f(v)$ ou loi de Erlang) est tracée sur le graphique de la planche C pour $k = 3$, 6, 9, 12 et 18 pièces lorsque le taux de panne est $\lambda = 3/4$. La courbe correspondant au cas $k = 12$ pièces est ombragée. La partie gris foncé, qui correspond aux cas où $v > 10$, représente environ 92,1% de la surface sous la courbe; le complément, 7,9%, est la probabilité théorique d'échec de la mission. Il est évident que si on s'était limité à $k = 3$ ou 6 pièces, les chances de succès auraient été trop minces (respectivement 2% et 24,1%). En revanche, on aurait été certain à 99,9% de réussir la mission en emportant $k = 18$ pièces. Le calcul des probabilités permet donc de quantifier le risque et de trouver un bon compromis au vu des contraintes de poids, d'espace, voire même de coût.

La planche D illustre la relation $\Pr(V > t) = \Pr(N \leq k - 1)$. Ce graphique peut être lu de deux manières complémentaires. L'abscisse s'étend de 0 à 10 ans ; l'ordonnée représente des probabilités, exprimées en pourcentages. Les plages de couleurs identifiées 1 à 12 sont associées aux durées de vie des pièces ; la partie blanche visible dans le coin supérieur droit représente un échec de la mission (le code de couleurs est le même que dans la planche A).



Pour lire la planche D, fixons-nous une verticale, disons au temps $t = 1$ an. La longueur du segment vertical vert représente alors la probabilité que la pièce n° 1 soit encore opérationnelle au temps t. La longueur du segment suivant est la probabilité que la pièce n° 2 soit en fonction à cet instant, et ainsi de suite pour les autres segments. Si on fait maintenant glisser t vers la droite, disons à 2 ans, on constate que le segment vert est plus petit. Ceci correspond au fait qu'à mesure que le temps passe, les chances de survie de la pièce n° 1 s'amenuisent ; de fait, elles diminuent à un taux exponentiel, comme on l'a vu précédemment.

De façon semblable, la frontière supérieure de la zone $i \in \{1, \ldots, 12\}$ représente la fonction de survie de la variable $V_1 + \cdots + V_i$, c'est-à-dire la courbe $\Pr(V_1 + \cdots + V_i > t)$ tracée en fonction de t. Comme nous l'avons vu, cette probabilité est aussi celle d'observer $N \leq i - 1$ pannes dans l'intervalle de temps $[0, t]$. Quand $i = 12$, cette probabilité est celle que la mission soit couronnée de succès ; comme on pourrait s'y attendre, elle décroît au fil du temps et est d'environ 92,1% au bout de 10 ans.

Les planches C et D sont fondées sur des calculs théoriques, tandis que les planches A et B représentent le fruit d'une simulation, répétée 25 fois. Si les postulats qui sous-tendent la théorie sont valables, les formules qui en découlent permettent de prédire le portrait qui se dégagerait si on effectuait un très grand nombre d'essais indépendants. C'est ainsi, par exemple, que la répartition des couleurs le long de la verticale $t = 10$ ans renseigne sur les chances que telle ou telle pièce soit en service au moment du rendez-vous entre la sonde et la comète. Quant à elle, la planche A' montre la répartition observée de façon empirique (on a simplement réordonné les résultats déjà rapportés à la planche A). Compte tenu du petit nombre de répétitions, la correspondance est plutôt bonne !