Promenades aléatoires ± 1

Gérard LETAC,
département Informatique de l'Institut Universitaire de Technologie de Clermont-Ferrand

Nous voudrions montrer ici, à travers quelques exemples, combien le modèle probabiliste d'une promenade aléatoire ± 1 est riche et comment il rend service dans des questions qui à première vue en paraissent fort éloignées.

Considérons un point P mobile sur la droite Z des entiers relatifs. A l'instant k (entier $\geqslant 0$), la position de P est repérée par l'entier S_k de Z. On suppose $S_0 = 0$. On jette une pièce de monnaie entre les instants k et $k+1$. Si c'est *face* on pose $S_{k+1} = S_k+1$, sinon $S_{k+1} = S_k-1$. En d'autres termes $S_k = X_1+\dots+X_k$, où $X_1, \dots, X_k$ sont des variables aléatoires indépendantes telles que $Pr(X_k = +1) = Pr(X_k = -1) = \frac{1}{2}$. Par exemple :

k	1	2	3	4	5	6	7
Pièce	pile	pile	face	face	face	pile	face
X_k	-1	-1	1	1	1	-1	1
S_k	-1	-2	-1	0	1	0	1

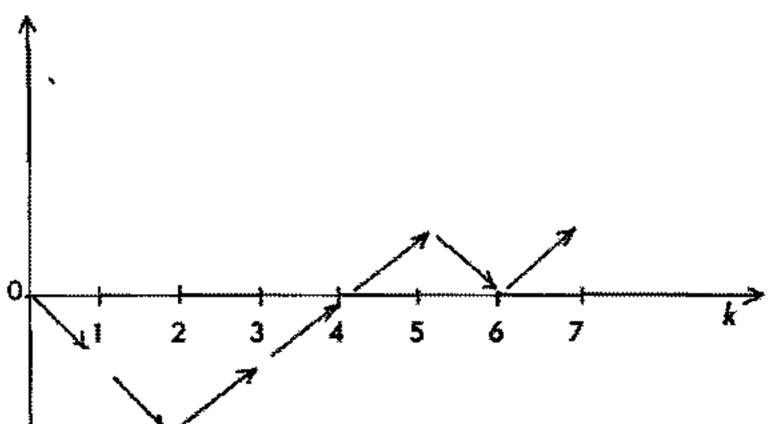

FIG. 1.

1. Une réussite.

Dans un jeu de 52 cartes, Pierre tire une carte et, avant de la regarder, devine sa couleur : rouge ou noire. Il marque un point s'il a deviné juste. Il recommence avec le paquet de 51 cartes restantes, et ainsi de suite jusqu'à épuisement du paquet. Paul lui donne 1 franc par point marqué. Combien d'argent Pierre doit-il donner à Paul au début du jeu pour que celui-ci soit équitable? Si vous préférez, quelle est la moyenne E(N) — ou espérance mathématique — du nombre N de points marqués?

Contrairement à ce qu'on pourrait penser à première vue, 26 n'est pas E(N) ; c'est même le salaire minimum garanti de Pierre : qu'il lui suffise d'annoncer 52 fois rouge. Il y a mieux : Pierre peut annoncer *rouge* si c'est la couleur la mieux représentée dans le paquet (ou vice versa) et annoncer au hasard s'il y a égalité. (Il n'est pas difficile de vérifier que cette méthode maximise E(N).)

Introduisons les variables aléatoires $Y_1, Y_2, \ldots, Y_{52}$ où $Y_k = +1$ si la $k^{\text{ème}}$ carte tirée est rouge et $Y_k = -1$ sinon, et posons $Z_k = Y_1+\ldots+Y_k$, $Z_0 = 0$. Le nombre de points obtenus est donc 26 plus le nombre de fois où Pierre a bien deviné quand il choisissait au hasard. Soit K le nombre de fois où il choisissait au hasard :

la moyenne du nombre de points est donc $26+\dfrac{1}{2}\,E(K)$.

Il est clair que K est égal au nombre de zéros dans $Z_0, Z_1, \ldots, Z_{51}$. Hélas les Y_k n'engendrent pas une promenade ± 1, car $Z_{52} = 0$ empêche l'indépendance. L'idée est pourtant de « plonger » le processus dans une promenade en utilisant les probabilités conditionnelles.

Considérons donc une « vraie » promenade $(S_k)\,k_{k=0}^{\infty} = 0$. On pose :

$$T_0 = 0, \qquad T_{n+1} = \inf\{k > T_n \text{ tels que } S_k = 0\}.$$

On voit que T_n est l'instant du $n^{\text{ième}}$ retour à l'origine de la promenade. Dans l'exemple de la figure 1, $T_1 = 4$ et $T_2 = 6$. Il est possible de montrer que T_n existe avec la probabilité 1. Si nous remarquons que $P(S_n = 0) = 0$ si n est impair et que $P(S_n = 0) = C_n^{n/2}\,\dfrac{1}{2^n}$ si n est pair (analyse combinatoire facile), on a l'identité suivante, obtenue grâce à la formule du binôme :

$$\begin{aligned}(1-t^2)^{-1/2} &= \sum_{n=0}^{\infty} t^n P(S_n = 0) \qquad = \sum_{n=0}^{\infty} t^n P\Big(\bigcup_{k=0}^{n} T_k = n\Big)\\ &= \sum_{n=0}^{\infty} t^n \sum_{k=0}^{n} P(T_k = n) = \sum_{k=0}^{\infty}\ \sum_{n=k}^{\infty} t^n P(T_k = n).\end{aligned}$$

Un peu d'intuition permet de penser que les longueurs des intervalles de temps séparant ces différents retours sont indépendants et de même loi que T_1. Ces intervalles étant $I_1 = T_1$, $I_n = T_n - T_{n-1}$, on a donc $T_n = I_1+\ldots+I_n$.

Par conséquent : $\sum_{n=k}^{\infty} t^n P(T_k = n) = \left[\sum_{n=1}^{\infty} t^n P(T_1 = n)\right]^k$.

D'où la jolie identité : $\sum_{n=1}^{\infty} t^n P(T_1 = n) = 1 - \sqrt{1-t^2}$.

Désignons par K_n le nombre de zéros de la suite $S_0, S_1, \ldots, S_{n-1}$.

Si on revient au problème posé, il est clair que $E(K_{52}|S_{52} = 0) = E(K)$. Comme il est fréquent en mathématiques, on a intérêt à résoudre un problème plus général : le calcul de $E(K_n|S_n = 0)$.

$$\text{Or : } \sum_{n=0}^{\infty} t^n P(S_n = 0).E(K_n|S_n = 0) = 1 + \sum_{n=1}^{\infty} t^n P(S_n = 0).\sum_{k=1}^{\infty} kP(K_n = k|S_n = 0)$$

$$= 1 + \sum_{k=1}^{\infty} k \sum_{n=k}^{\infty} t^n P(K_n = k \cap S_n = 0).$$

Le point essentiel est que $P(K_n = k \cap S_n = 0) = P(T_k = n)$.

Les expressions ci-dessus sont donc égales à

$$1 + \sum_{k=1}^{\infty} k(1-\sqrt{1-t^2})^k = 1 + \frac{1}{1-t^2} - \frac{1}{\sqrt{1-t^2}}$$

Pour avoir $E(K_{2n}|S_{2n} = 0)$, il suffit d'examiner les coefficients de t^{2n} et on obtient $2^{2n}/C_{2n}^n - 1$, soit, si $2n = 52$, environ 8, 03. Donc $E(N) \# 30, 02$.

Précisons qu'on peut perfectionner la technique pour montrer que la variance de N est 2,41.

2. Points au hasard sur un cercle.

Un problème qui intéresse le statisticien est le suivant : sur un cercle $\mathcal{T}$, choisissons « au hasard » n points $P_1, \ldots, P_n$ (sous-entendu : uniformément et indépendamment); on fait passer par le centre O une droite D de sorte que le nombre N des P_i situés d'un même côté de la droite soit le plus grand possible. Il est clair que $2N \geqslant n$. Mais quelle est la loi de N? Autrement dit : calculer $Pr\ (N = k)$, si $\frac{n}{2} \leqslant k \leqslant n$.

Sur $\mathcal{T}$, plaçons les points $P'_1, \ldots, P'_n$ définis par $\overrightarrow{OP_i} + \overrightarrow{OP'_i} = \vec{0}$. Fixons une origine I et un sens de parcours sur $\mathcal{T}$ (fig. 2). Les $2n$ points $P_1 \ldots P_n, P'_1, \ldots, P'_n$ sont donc classés par ce sens de parcours. On définit les variables aléatoires $X_1, \ldots, X_{2n}$ par $X_k = +1$ si le $k^{\text{ème}}$ point rencontré est un P, $X_k = -1$ si c'est un P'. Évidemment, par suite des hypothèses, $X_1, \ldots, X_n$ sont indépendants et tels que $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$. Et $X_{i+n} + X_i = 0$ si $i \in [1, n]$.

On pose $S_k = X_1 + \ldots + X_k$.

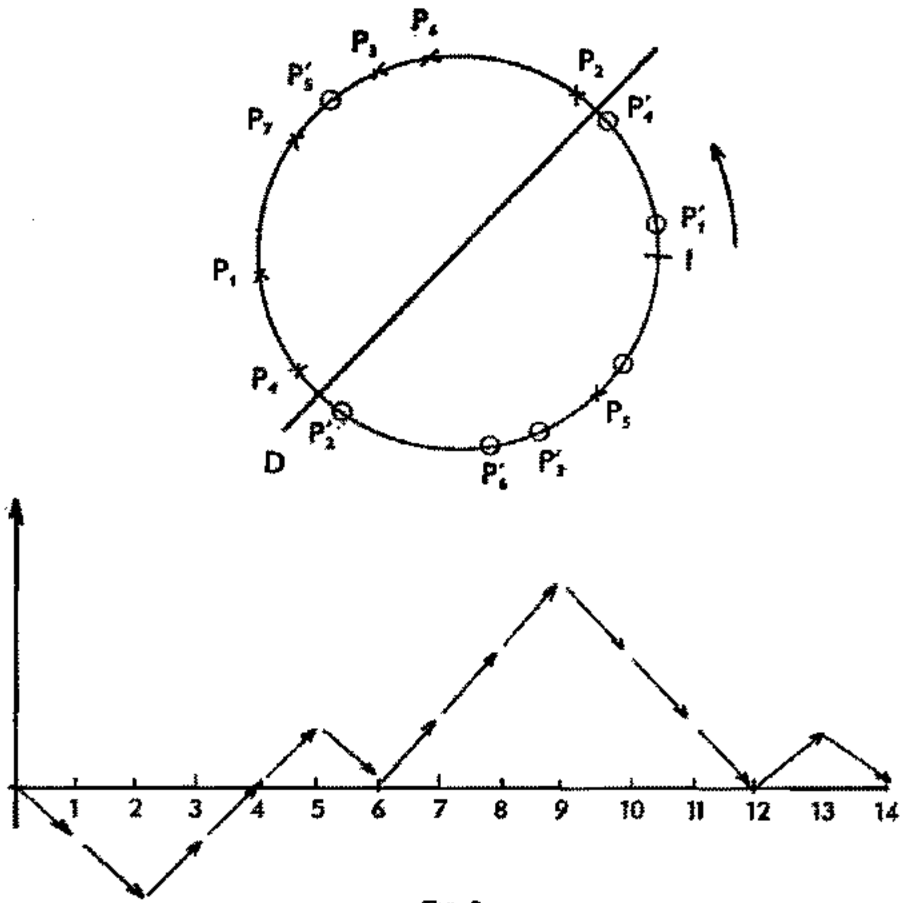

FIG. 2.

J'énonce maintenant un lemme que le lecteur patient n'aura pas de difficulté à démontrer :

Lemme. Soient $\varepsilon_1, \ldots, \varepsilon_{2n}$ des nombres égaux à ± 1 tels que

$$\varepsilon_{i+n} + \varepsilon_i = 0 \quad \text{si} \quad i \in [1, n].$$

On pose

$$\varepsilon_i^+ = \max(0, \varepsilon_i), \qquad f_k = \varepsilon_1 + \ldots + \varepsilon_k \qquad \text{et} \qquad g_k = \varepsilon_1^+ + \ldots + \varepsilon_k^+.$$

Alors $$\max_{1 \leqslant k \leqslant 2n} f_k - \min_{1 \leqslant k \leqslant 2n} f_k = 2 \max_{1 \leqslant k \leqslant n} (g_{k+n} - g_k) - n.$$

Un peu de réflexion — et l'examen de la figure 2 — indique la signification du lemme : l'oscillation de S_k entre 1 et $2n$ est égale à $2N - n$. Les techniques de promenades aléatoires ± 1, trop longues pour être développées ici, permettent d'obtenir la loi de l'oscillation $O_n = \max_{1 \leqslant k \leqslant 2n} S_k - \min_{1 \leqslant k \leqslant 2n} S_k$.

Un cas est particulièrement simple, celui où $N = n$. Alors $O_n = n$, ce qui ne peut arriver que s'il existe i, entre 1 et n, tel que

$$X_1 = X_2 = \ldots = X_i = -X_{i+1} = \ldots = -X_n$$

comme le montrerait le dessin. La probabilité est facile à calculer : c'est $\frac{n}{2^{n-1}}$.

3. *Comparaison de lois empiriques.*

Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et de même loi continue définie par $F(x) = P(Y_i < x)$. $F(x)$ n'est pas nécessairement connue, et pour avoir une idée de cette loi, le statisticien forme la fonction aléatoire suivante, appelée « loi empirique ».

$$F_n(x) = \frac{1}{n}(\text{nombre de } i \in [1, n) \text{ tels que } Y_i < x),$$

dont on peut démontrer qu'elle converge vers $F(x)$ si $n \to \infty$.

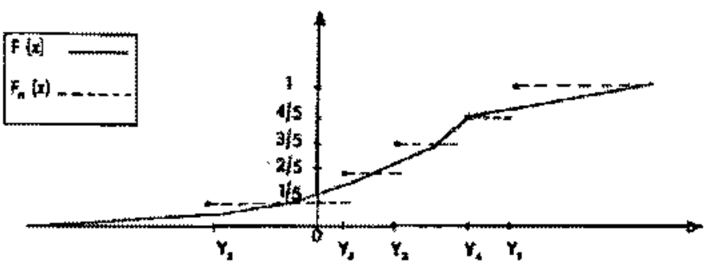

Fig. 3.

Soient ensuite $Z_1, \ldots, Z_n$ des variables aléatoires indépendantes et de même loi, dont on pense que la loi commune est F. Pour tester cette affirmation, rien de plus naturel que de regarder le maximum D_n des différences entre les lois empiriques. $F_n(x)$ et $G_n(x)$ des Y et Z :

$$D_n = \max_x [F_n(x) - G_n(x)].$$

Mais pour tester statistiquement l'hypothèse de l'identité des lois des Y et Z il est indispensable de connaître la loi de D_n dans le cas où cette hypothèse est vérifiée. Il est remarquable que dans ce cas non seulement la loi de D_n ne dépend pas de F, mais est reliée à de simples propriétés d'une promenade ± 1.

Classons les $2n$ nombres $Y_1, \ldots, Y_n, Z_1, \ldots, Z_n$ par ordre croissant et définissons les variables aléatoires $X_1, \ldots, X_{2n}$ par $X_k = +1$ si le $k^{\text{ième}}$ nombre est un Y et par $X_k = -1$ si c'est un Z.

Posons $S_k = X_1 + \ldots + X_k$. On voit alors que les k premiers termes de la suite croissante comprennent $\frac{1}{2}(k+S_k)$ nombres Y et $\frac{1}{2}(k-S_k)$ nombres Z ce qui entraîne l'existence d'un x tel que

$$F_n(x) - G_n(x) = \frac{S_n}{n}.$$

Les X_k ne sont pas indépendants puisque soumis à la restriction $S_{2n} = 0$. Cependant, on peut utiliser le même artifice que dans l'exemple 1, en plongeant les S_k dans une promenade ± 1 et en conditionnant par $S_{2n} = 0$. Supposons donc que les S_k forment une promenade ± 1, on a alors

$$P\left(D_n > \frac{d}{n}\right) = P(\exists k \in \{0, \ldots, n\} \text{ tel que } S_k > d;\ S_{2n} = 0).$$

Le calcul du membre de droite se fait en utilisant le célèbre « principe de symétrie » de Désiré André : dans un diagramme comme celui de la figure 1, il y a C_{2n}^n « chemins » équiprobables joignant les points (0, 0) et (2n, 0).

Le problème est de compter le nombre de ces chemins qui touchent la droite $y = d$. La figure 4 montre qu'il y a correspondance bijective entre ces derniers et les chemins joignant les points (0, 2d) et (2n, 0), cette correspondance étant réalisée en prenant le symétrique par rapport à $y = d$ de la partie du chemin de (0, 0) à (2n, 0) située entre (0, 0) et le premier point de contact avec $y = d$.

FIG. 4.

La probabilité cherchée est donc

$$C_{2n}^{n-d} / C_{2n}^{n} = \frac{n!}{(n-d)!} \cdot \frac{n!}{(n+d)!}$$

Bibliographie

[1] LETAC (G.). — *Problèmes de Probabilité*, P.U.F., à paraître.
[2] HODGES (J. L.). (1955). — A bivariate sign test. *Ann. Math. Statist.*, 26, 523-527.
[3] GNEDENKO et KOROLJUK (1951).