La prédiction des longues suites de succès

Jean-François Kentzel [*],
Thierry de la Rue [†]

Résumé

Nous nous intéressons à une expérience aléatoire consistant à effectuer successivement n tirages à pile ou face, et à regarder la plus longue suite de faces consécutives obtenues. De manière assez incroyable, même lorsque n devient extrêmement grand, la longueur de cette plus longue suite de faces peut être prédite à une ou deux unités près avec une probabilité très proche de 1 !

Nous donnons quelques arguments mathématiques qui expliquent ce phénomène.

1 Quelques expériences

La figure 1 a été réalisée en simulant 20 suites de 100 tirages à pile ou face consécutifs, représentées sur chacune des 20 lignes de la figure (les piles sont en blanc). Sur chaque ligne on a recherché la plus longue suite de faces, et on a recopié cette suite à droite pour mieux visualiser sa longueur.

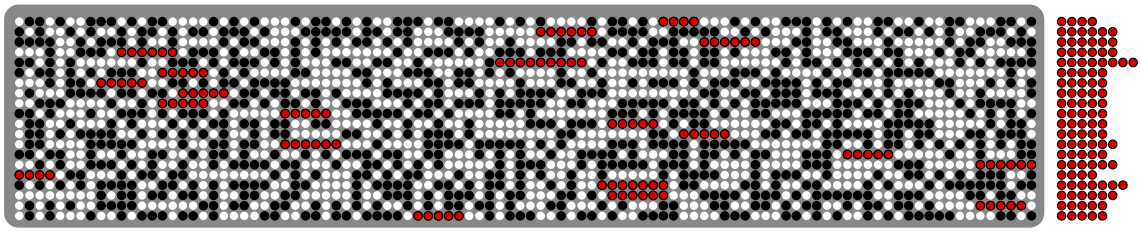

FIGURE 1 – Visualisation des plus longues suites de faces sur 100 tirages successifs

Le bilan de cette petite simulation montre une nette prédominance des longueurs 5 et 6. En répétant l'expérience un plus grand nombre de fois, toujours avec 100 tirages consécutifs, on observe que la plus grande suite de faces se répartit suivant les fréquences suivantes.

Longueur maximum	3	4	5	6	7	8	9	10	11
Fréquence observée	2%	15%	26%	23%	15%	8%	4%	2%	1%

Ainsi, bien que tous les résultats entre 0 et 100 soient possibles, plus de 95% des expériences aboutissent à une longueur maximum entre 3 et 11. Et si on devait parier à l'avance sur la plus grande longueur de faces observée, on voit qu'en misant sur « 5 » on aurait un peu plus d'une chance sur 4 de gagner.

[*]Lycée Pardailhan, Auch
[†]Laboratoire de Mathématiques Raphaël Salem, Rouen

Effectuons maintenant la même expérience, mais cette fois avec un nombre de lancers successifs beaucoup, beaucoup plus grand. Disons un million de fois plus grand. La plus longue suite de faces successives peut alors être de n'importe quelle longueur entre 0 et 100 000 000, et on s'attend plutôt à ce que cette longueur soit très difficile à prévoir. Pourtant, voici les fréquences mesurées en répétant 1 000 fois l'expérience :

Longueur maximum	23	24	25	26	27	28	29	30	31
Fréquence observée	2%	14%	24%	17%	16%	8%	7%	5%	7%

De manière incroyable, la longueur de la plus grande succession de faces en 100 millions de lancers n'est pas plus difficile à prévoir que lorsqu'on effectuait seulement 100 lancers ! En fait elle apparaît même plus prévisible, puisqu'on constate que 100% des expériences ont donné une suite de longueur comprise entre 23 et 31. On voit également qu'en pariant sur une longueur maximum de 25 faces consécutifs, on a environ une chance sur quatre de gagner !

On pourrait suspecter que ces résultats étranges sont dus à l'imperfection des simulations effectuées sur l'ordinateur, et que le générateur de nombres aléatoires utilisé est mis en défaut par ce test. Pourtant, nous allons voir dans la suite de l'article qu'il n'en est rien : l'incroyable prévisibilité de cette plus longue suite de faces consécutifs peut être expliquée mathématiquement. Pour une pièce équilibrée, nous allons notamment montrer que, si le nombre de lancers est assez grand, il existe toujours un entier k tel que la plus grande suite de faces soit de longueur k avec une probabilité d'au moins $0,236$.

2 Formalisation mathématique

Nous allons modéliser l'expérience décrite ci-dessus par une suite $(B_j)_{1\leq j\leq n}$ de variables aléatoires de Bernoulli indépendantes et de même loi, l'entier n correspondant au nombre de lancers de l'expérience. Chaque B_j peut donc prendre les valeurs 0 (échec, ou pile) ou 1 (succès, ou face), et nous noterons $p \in]0,1[$ la probabilité d'un succès. A priori p n'est pas nécessairement égale à $1/2$: comme nous allons le voir le phénomène existe même pour des pièces non équilibrées (et en fait il est d'autant plus marqué lorsque p est petit). On construit la variable aléatoire L_n comme la longueur maximum d'une suite de succès consécutifs observée au moins une fois dans la suite $(B_j)_{1\leq j\leq n}$. Nous nous intéressons à la loi de L_n, dont nous allons montrer qu'elle reste essentiellement concentrée sur un petit nombre d'entiers même quand n devient grand.

Cette variable aléatoire L_n est connue dans la littérature sous le nom de « longest head run » (voir notamment [4, 5, 6]).

2.1 Une relation de récurrence pour la loi de L_n

L'entier $k \geq 1$ étant fixé, on se propose dans un premier temps d'établir une relation de récurrence donnant $\mathbb{P}(L_n \geq k)$ pour tout $n \geq 0$. (Pour simplifier la formulation de la récurrence, on définit de manière naturelle L_0 comme étant constante, égale à 0.) On a évidemment $\mathbb{P}(L_n \geq k) = 0$ si $0 \leq n < k$, et

$$\mathbb{P}(L_k \geq k) = p^k.$$

Puis, pour $n \geq k+1$ on remarque que l'événement $\{L_n \geq k\}$ se traduit par

– soit il y avait déjà au moins une suite de k succès dans les $(n-1)$ premiers lancers,

– soit il n'y avait pas de telle suite dans les $n-k-1$ premiers lancers, et une suite d'exactement k succès consécutifs apparaît à la fin des n lancers sous la forme

$$\dots 0 \underbrace{1 \dots 1}_{k \text{ succès}} .$$

On obtient ainsi la formule de récurrence :

$$\mathbb{P}\left(L_n \geq k\right) = \mathbb{P}\left(L_{n-1} \geq k\right) + \Big(1 - \mathbb{P}\left(L_{n-k-1} \geq k\right)\Big)(1-p\,)p^k. \tag{1}$$

Cette relation de récurrence, avec les conditions initiales évidentes, définit totalement la loi de la variable aléatoire L_n, mais ne permet pas à notre connaissance d'en donner des formules facilement exploitables pour $\mathbb{P}(L_n = k)$. Elle est toutefois très facile à implémenter informatiquement, et nous pouvons ainsi par exemple calculer les premières valeurs (en n et en k) de $\mathbb{P}(L_n = k)$. Nous visualisons ces premières valeurs sur la figure 2 sous forme de graphes de $\mathbb{P}(L_n = k)$ en fonction de n qui varie entre 1 et 100. (Chaque graphe correspondant à une valeur différente de k, le paramètre p étant fixé à $1/2$).

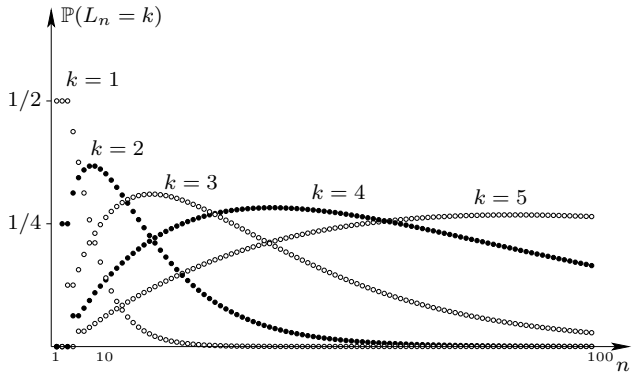

FIGURE 2 – Tracé de $\mathbb{P}(L_n = k)$ en fonction de n, pour plusieurs valeurs de k. Ici p vaut $1/2$.

On constate sur la figure que, pour chacune de ces premières valeurs de n, il existe toujours un entier k tel que $\mathbb{P}(L_n = k)$ atteigne une valeur au moins de l'ordre de $1/4$. Notre but dans la suite est de montrer que ce phénomène perdure lorsque n va à l'infini. Nous aurons besoin pour cela d'une approximation raisonnable de $\mathbb{P}(L_n = k)$, qui est l'objet de la section suivante.

3 Estimations de $\mathbb{P}(L_n = k)$

3.1 Approximation à l'aide d'une loi de Poisson

Fixons les entiers $1 \leq k \leq n-1$, et cherchons à estimer $\mathbb{P}(L_n \geq k)$ lorsque k et n/k sont très grands. Pour chaque j entre 0 et $n-k$, définissons la variable aléatoire S_j valant 1 si une suite d'au moins k succès commence exactement en $j+1$, valant 0 sinon : ainsi S_0 peut être vue comme le produit $B_1 \cdots B_k$, et pour $1 \leq j \leq n-k$, $S_j = (1-B_j)B_{j+1} \cdots B_{j+k}$. Les S_j sont donc aussi des variables de Bernoulli, de moyenne p^k pour $j = 0$ et $(1-p)p^k$ pour $1 \leq j \leq n-k$. Considérons la somme $W \overset{\text{déf}}{=} S_0 + S_1 + \cdots + S_{n-k}$: on a alors l'égalité des événements

$$\{W = 0\} = \{L_n < k\}. \tag{2}$$

Rappelons qu'une variable aléatoire Z à valeurs dans les entiers positifs est de Poisson de paramètre λ si, pour tout entier $r \geq 0$,

$$\mathbb{P}(Z = r) = e^{-\lambda}\frac{\lambda^r}{r!}.$$

Selon le paradigme de Poisson, que l'on peut voir par exemple comme une conséquence de l'inégalité dite *de Le Cam* [1], cette loi de Poisson est une bonne approximation de la loi d'une somme d'un grand nombre de variables de Bernoulli indépendantes et d'espérances petites (le paramètre λ à utiliser étant la somme de ces espérances). Or, lorsque k est grand, chaque S_j est justement d'espérance petite. Si les S_j étaient indépendantes, on pourrait immédiatement approcher la loi de leur somme W par celle d'une variable aléatoire de Poisson de paramètre

$$\lambda \stackrel{\text{déf}}{=} \sum_{j=0}^{n-k} \mathbb{E}[S_j] = p^k\big(1 + (n-k)(1-p)\big). \tag{3}$$

Malheureusement, dans notre cas les variables S_j ne sont pas toutes indépendantes. Elles vérifient :

$$\text{si } |j_1 - j_2| > k,\ S_{j_1} \text{ et } S_{j_2} \text{ sont indépendantes,} \tag{4}$$

mais

$$\text{si } |j_1 - j_2| \leq k,\ S_{j_1}S_{j_2} = 0, \tag{5}$$

puisque deux suites de succès de longueur au moins k ne peuvent pas commencer à une distance inférieure ou égale à k. Néanmoins, l'approximation de la loi de W par une loi de Poisson de paramètre λ reste possible dans ce cadre : elle est prouvée par des résultats obtenus par la méthode dite de Stein-Chen. Leur preuve dépasse le cadre du présent article, nous nous contenterons d'énoncer le théorème suivant, adaptation du Théorème 1 de [3] à notre situation.

Théorème 3.1. *Soient $S_0, \ldots, S_{n-k}$ des variables de Bernoulli satisfaisant* (4) *et* (5). *Soit Z une variable aléatoire de Poisson de paramètre λ égal à l'espérance de la somme W des S_j. Alors*

$$\sum_{r\geq 0} |\mathbb{P}(W = r) - \mathbb{P}(Z = r)| \leq b/\lambda,$$

où

$$b \stackrel{\text{déf}}{=} \sum_{|j_1 - j_2| \leq k} \mathbb{P}(S_{j_2} = 1)\,\mathbb{P}(S_{j_1} = 1).$$

3.2 L'estimation qui nous intéresse

Nous voulons maintenant appliquer le théorème précédent pour donner une estimation de $\mathbb{P}(L_n < k) = \mathbb{P}(W = 0)$. Dans notre situation, puisque $\mathbb{P}(S_j = 1) \leq p^k$ pour tout j, nous pouvons majorer la quantité b par $p^{2k}(n-k+1)(2k+1)$. Par ailleurs, nous déduisons de (3) que

$$\lambda \geq p^k(1-p)(n-k+1).$$

En utilisant seulement le terme correspondant à $r = 0$ dans l'inégalité donnée par le Théorème 3.1, on en déduit

$$|\mathbb{P}(L_n < k) - \exp(-\lambda)| \leq \frac{p^k(2k+1)}{1-p}. \tag{6}$$

Puis, pour faciliter les calculs, remplaçons dans l'estimation ci-dessus la vraie valeur de λ par l'expression plus simple $p^k n(1-p)$ (ce qui revient à négliger les « effets de bords »). Ce faisant, l'erreur commise est au plus

$$\left|\exp(-\lambda) - \exp\left(-p^k n(1-p)\right)\right| \leq \left|\lambda - p^k n(1-p)\right| \leq (k+1)p^k. \tag{7}$$

En utilisant finalement le fait que $\mathbb{P}(L_n = k) = \mathbb{P}(L_n < k+1) - \mathbb{P}(L_n < k)$, nous obtenons en combinant (6) et (7) pour k et $k+1$ le résultat suivant.

Proposition 3.2. *Pour tous* $1 \leq k < n$,

$$\left|\mathbb{P}(L_n = k) - f_p\left(p^k n(1-p)\right)\right| \leq p^k \left(\frac{k(1+p)(3-p)+3p+1}{1-p} + 2p + 1\right), \tag{8}$$

où la fonction f_p *est définie par*

$$f_p(x) \stackrel{\text{déf}}{=} e^{-px} - e^{-x}.$$

3.3 Le mode de L_n

Une analyse élémentaire des variations de f_p sur $\mathbb{R}_+$ montre que f_p est croissante jusqu'à son unique maximum, puis décroissante en tendant vers 0 à l'infini. Il est facile d'en déduire l'existence d'un unique point $x_p > 0$ vérifiant $f_p(x_p) = f_p(px_p)$. On définit alors $m_p \stackrel{\text{déf}}{=} f_p(x_p)$ (voir Figure 3).

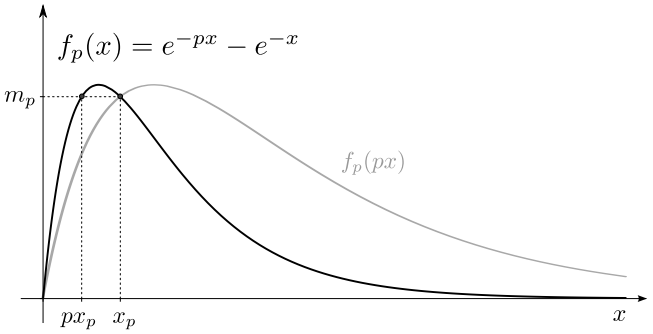

FIGURE 3 – graphe de la fonction f_p (pour $p = 1/2$)

Dès que n est assez grand pour que $n(1-p)$ dépasse x_p, il existe un unique entier $k(n) \geq 1$ tel que $px_p \leq p^{k(n)} n(1-p) < x_p$: $k(n)$ est donné par l'expression

$$k(n) \stackrel{\text{déf}}{=} \left[\frac{\ln n}{\ln 1/p} + \frac{\ln \frac{1-p}{x_p}}{\ln 1/p} + 1\right],$$

où $[\,\cdot\,]$ désigne la partie entière. Cet entier $k(n)$ réalise le maximum sur tous les $k \geq 1$ de $f_p\left(p^k n(1-p)\right)$: il vérifie

$$f_p\left(p^{k(n)} n(1-p)\right) \geq m_p.$$

On voit que, quand $n \to \infty$, $k(n)$ est équivalent à $\ln n / \ln(1/p)$. En particulier $k(n)$ est nettement plus petit que n, mais tend vers l'infini. Lorsque $k = k(n)$, le membre de droite dans l'estimation (8) devient donc négligeable quand $n \to \infty$, et on en déduit

$$\liminf_{n\to\infty} \max_{k\geq 1} \mathbb{P}(L_n = k) \geq \liminf_{n\to\infty} \mathbb{P}\big(L_n = k(n)\big) \geq m_p. \tag{9}$$

Il est naturel alors de chercher à en savoir un peu plus sur la quantité m_p. Malheureusement nous ne disposons pas d'expression explicite de cette quantité, sauf pour quelques valeurs de p. Dans le cas équilibré $p = 1/2$ notamment, il est facile de voir que la détermination de $x_{1/2}$ se ramène à la résolution d'une équation du second degré : en effet, $x_{1/2}$ vérifie

$$f_{1/2}\left(x_{1/2}\right) = f_{1/2}\left(\frac{x_{1/2}}{2}\right).$$

En posant $z \overset{\text{déf}}{=} e^{-\frac{x_{1/2}}{4}}$, l'égalité ci-dessus se traduit par

$$z - z^2 = z^2 - z^4,$$

autrement dit z est racine du polynôme

$$Z^4 - 2Z^2 + Z = Z(Z-1)(Z^2 + Z - 1).$$

Comme z est strictement compris entre 0 et 1, on en déduit que $z = \frac{1}{2}(\sqrt{5}-1)$ (c'est l'inverse du nombre d'or !), puis que

$$m_{1/2} = f_{1/2}\left(\frac{x_{1/2}}{2}\right) = z - z^2 = \sqrt{5} - 2 \simeq 0,236.$$

La figure 4 montre le graphe de m_p en fonction de p, obtenu en résolvant numériquement l'équation $f_p(y) = f_p(py)$. La fonction $p \mapsto m_p$ y apparaît décroissante, mais nous n'avons pas de démonstration complète de cette propriété.

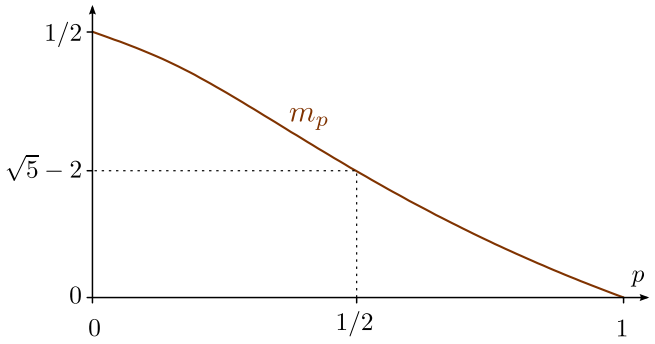

FIGURE 4 – Graphe de m_p en fonction de p .

On peut cependant montrer que m_p tend vers $1/2$ quand $p \to 0$. Ainsi, pour p assez petit et n assez grand, il existe un entier k tel que la plus longue suite de succès en n épreuves de Bernoulli soit de longueur k avec pratiquement une chance sur deux !

4 Généralisation à d'autres variables aléatoires du même type

Plus généralement, pour un entier $r \geq 1$ fixé, on peut aussi étudier la variable aléatoire L_n^r qui vaut la longueur maximum d'une suite de succès consécutifs observée au moins r fois dans la suite $(B_j)_{1\leq j\leq n}$, chacune des r occurences étant séparée des autres (si $r > 1$) par au moins un échec. Avec ces notations, on a $L_n = L_n^1$. Par exemple, pour $n = 20$, si la suite observée est

$$1\,1\,1\,1\;0\;1\,1\,1\;0\;1\,1\;0\;1\,1\,1\,1\;0\;1\,1\,1,$$

on obtient $L_{20} = L_{20}^2 = 4$, $L_{20}^3 = L_{20}^4 = 3$, $L_{20}^5 = 2$ et $L_{20}^r = 0$ si $r > 5$.

La même méthode que celle présentée ici s'applique pour chaque variable L_n^r et fournit des résultats similaires, avec des probabilités encore plus grandes lorsque $r > 1$. Le lecteur intéressé pourra trouver des détails dans les notes du premier auteur, lisibles en ligne [2].

Références

[1] *Inégalité de Le Cam*, fr.wikipedia.org/wiki/Inégalité_de_Le_Cam.

[2] *Notes de J.F. Kentzel en ligne*, pardailhan.entmip.fr/rubrique-des-disciplines/mathematiques/documents-enseignants/doc-k-3359.htm.

[3] R. Arratia, L. Goldstein, et L. Gordon, *Two moments suffice for Poisson approximations : the Chen-Stein method*, Ann. Probab. **17** (1989), no. 1, 9–25.

[4] Louis Gordon, Mark F. Schilling, et Michael S. Waterman, *An extreme value theory for long head runs*, Probab. Theory Relat. Fields **72** (1986), no. 2, 279–287.

[5] Mark F. Schilling, *The longest run of heads*, College Math. J. **21** (1990), no. 3, 196–207.

[6] Mark F. Schilling, *The surprising predictability of long runs*, Math. Mag. **85** (2012), no. 2, 141–149.