Partiel de Probabilités corrigé

1 Jeux de pile/face non transitif

Problème 1 : Motivation biologique : Pour détecter les séquences génétiques sous sélection (donc sûrement fonctionnelles), des méthodes statistiques se sont intéressées à étudier leurs fréquences et les relation entre elles.

Pour simplifier l'analyse, nous allons considérer une séquence de pile ou face symétrique : au n^e tirage, on tire P ou F avec probabilité 0,5 chacun et les tirages sont indépendants. On va voir que les notions de " fréquence d'apparition ", de " temps moyen d'attente " et de " dominance " ne sont pas aussi identiques qu'on pourrait le penser.

1) Quelle est la probabilité d'avoir la séquence PFPF à une position donnée ?

*Reponse :*PFPF apparaît à la position $n+3$ si et seulement si aux positions n, $n+1$, $n+2$ et $n+3$, les tirages sont respectivement P, F, P et F. Par indépendance des tirages, la probabilité du motif $PFPF$ est le produit des 4 probabilités pour chacun des tirages, soit 1/16.

2) En utilisant la linéarité de l'espérance, calculer le nombre moyen d'occurrences de cette séquence PFFP parmi 203 tirages de pile/face.

Indication : on pensera à décomposer ce nombre d'occurrences selon la position où celle-ci pourrait se produire.

*Reponse :*On peut écrire le nombre d'occurrence comme la somme de variables aléatoires indexées par n qui indique si le motif est présent en position n. Les variables sont toutes de Bernoulli de paramètre 1/16. Seules les positions plus grandes que 4 peuvent avoir un motif. Pour calculer la moyenne, on utilise la linéarité de la moyenne. Ce nombre moyen d'occurrences vaut donc le produit des moyennes de Bernoulli, à savoir 1/16, par le nombre de telles Bernoulli, à savoir 200. On obtient ainsi une moyenne de 12,5 occurrences.

3) En déduire que toutes les séquences de même longueur (ici 4 pour PFPF) ont la même fréquence d'apparitions.

Reponse :: Tous les motifs de même longueur ont la même probabilité d'apparaître vu le raisonnement de la question 1. Le nombre moyen d'occurrence calculé à la question précédente est donc le même, et ce même si on changeait la longueur de la séquence considérée. Or, par la Loi des Grands Nombres, la fréquence d'occurrences doit se stabiliser lorsque la séquence est très grande vers une fréquence asymptotique (désignée comme fréquence d'apparitions), qui est la valeur moyenne recherchée. Cette valeur ne dépend donc pas du motif, à longueur fixée.

On cherche ici à comparer le nombre moyen de tirages nécessaires entre deux séquences successives de PF versus celui entre deux séquences de FF et de l'autre FF. On les notera respectivement T_1 et T_2, et on admettra que ces quantités sont finies.

4) Notons $T_1(F)$ le nombre moyen de tirages supplémentaires avant d'atteindre PF sachant que le premier tirage est face. De même pour $T_1(P)$ où le premier tirage est pile. Avec la loi du deuxième tirage, trouver deux relation entre ces deux quantités.

*Reponse :*Soit A_1 la variable aléatoire du nombre de tirages avant la première occurrence de PF avec F déjà présent. Soit de même B_1 en supposant que P est déjà présent. Si le tirage suivant T_2 est F (i.e. $T_2 = F$), on ne peut pas avoir un motif PF et le fait que T_1 vaux F n'a plus d'effet sur l'apparition d'un nouveau PF. La variable d'attente avant le prochain PF a alors la même loi que A_1. Cela signifie que A_1 conditionné à ce que $\{T_2 = F\}$ a la même loi que $1 + A_1$ sans conditionnement.

De même si le 2e tirage est P : A_1 conditionné à ce que $\{T_2 = P\}$ a la même loi que $1 + B_1$ sans conditionnement.

Chaque événement sur le premier tirage arrive avec probabilité 1/2. En prenant la moyenne, on en déduit que $T_1(F) = (1/2) \times (1 + T_1(F)) + (1/2) \times (1 + T_1(P)) = 1 + (T_1(F) + T_1(P))/2$.

Le raisonnement partant de P est similaire, si ce n'est que l'on aboutit au motif dès lors que F est sorti au premier tirage. On en déduit : $T_1(P) = (1/2) \times 1 + (1/2) \times (1 + T_1(P)) = 1 + T_1(P)/2$.

5) En déduire l'expression de T_1.

*Reponse :*On observe immédiatement que $T_1 = T_1(F)$. L'équation sur $T_1(P)$ conduit à $T_1(P) = 2$. Via celle sur $T_1(F)$, on conclut ainsi $T_1 = T_1(F) = 4$.

6) Faire de même pour T_2. Obtient-on la même valeur ? Comment l'expliquer vu 3) ?

*Reponse :*Par symétrie, les expressions pour $T_2(P)$ et $T_2(F)$ sont exactement celles de $T_1(F)$ et $T_2(P)$ (en renversant les rôles des premières entrées). Néanmoins, on a cette fois-ci que $T_2 = T_2(F) = 2$. On ne trouve pas les mêmes valeurs car les chevauchements introduisent des corrélations entre les position d'occurrences.

On veut maintenant calculer la probabilité de trouver PF avant FF. On notera p_1 la probabilité de voir gagner PF en commençant par pile et p_2 en commençant par face.

7) De même que pour la question 4, trouver des relations entre ces deux quantités.

Reponse :: En adaptant le raisonnement de la question 4, on trouve en faisant la moyenne sur le deuxième tirage : $p_1 = 1/2 + 1/2 \times p_1$ et $p_2 = (p_1)/2$. Cela conduit immédiatement à $p_1 = 1$ et $p_2 = 1/2$.

Le fait que $p_1 = 1$ a un argument déterministe : on ne peut pas obtenir FF sans que le premier F de la séquence ne crée un motif PF !

8) En déduire la probabilité demandée. Est-elle cohérente avec le résultat en 6) ? *Reponse :*En faisant la moyenne sur le premier tirage, la probabilité demandée vaut $(p_1 + p_2)/2 = 3/4$. Cela peut paraître étonnant sachant que les deux motifs ont la même fréquence d'occurrence (cf. Q3). Cependant, comme on l'a vu à la question 6, les positions de ces occurrences sont couplées : les motifs FF ont tendance à se suivre, contrairement à PF. Puisqu'ils sont regroupés avec la même fréquence, les trains de F successifs doivent être plus éloignés les uns des autres.

2 Enjeu écologique du taux de croissance

Problème 2 : Michel et Béatrice se disputent sur la définition naturelle du taux de croissance d'une population. Dans leurs modèles élémentaires, ils décrivent l'évolution au cours des générations de la taille de population de poissons. Notons N_n cette taille à la génération n. Celle la génération suivante est donnée par $N_{n+1} = X_{n+1} \times N_n$, où le facteur multiplicatif X_{n+1} reflète l'aléa de l'environnement.

1) Ecrire N_n en fonction des X_i et de N_0. Dans le cas où les X_i sont constants à x, donner la valeur du taux de croissance r tel que N_n évolue comme e^{rn}.

*Reponse :*En écrivant N_n pour les premières valeurs de n, on trouve : $N_1 = X_1 N_0$, $N_2 = X_2 N_1 = X_2 X_1 N_0$. Par récurrence (évidente), on trouve $N_n = N_0 \times \prod_{i=1}^{n} X_i$

Le cas $x < 0$ n'est pas réaliste pour une taille de population et $x = 0$ guère intéressant. Si les X_i valent tous $x > 0$, on en déduit que

$$N_n = N_0 \times x^n = N_0 \times \exp[n \log(x)].$$

Le taux de croissance est donc $r = \log(x)$ dans ce cadre.

2) On veut dire qu'une espèce est en danger si son taux de croissance est négatif. A quoi cela correspond-il à la question précédente ?

Reponse :$r = \log(x) < 0$ correspond à $x \in (0, 1)$. A chaque étape, la taille de population est contracté par le facteur x et tend donc vers 0 à vitesse exponentielle. r quantifie l'échelle de temps du déclin. Ainsi, la population est divisée par deux à partir du moment où $n \geq \log(2)/(-r)$.

3) Michel défend l'idée de choisir $r_M := \log \mathbb{E}(X)$ comme définition du taux de croissance, tandis que Béatrice juge plus raisonnable de $r_B := \mathbb{E} \log(X)$. Par une inégalité classique, montrer qu'il y a plus d'espèces en danger selon la définition de Béatrice.

*Reponse :*La fonction : $x \mapsto -\log(x)$ est convexe, car de dérivée seconde : $x \mapsto x^{-2}$ positive. Par l'inégalité de Jensen, on en déduit que $\mathbb{E} - \log(X) \geq -\log \mathbb{E}(X)$, d'où $r_B \leq r_M$. $r_M < 0$ implique donc $r_B < 0$, ce qui signifie que le critère de Béatrice amène à considérer plus d'espèces comme en danger.

4) Calculer l'espérance de N_n. Est-ce que ce résultat va dans le sens de Michel ou de Béatrice ?

*Reponse :*Puisque les X_i sont indépendants, vu la formule donnée en Q1 :

$$\mathbb{E}(N_n) = \mathbb{E}(N_0 \times \prod_{i=1}^{n} X_i) = N_0 \times \prod_{i=1}^{n} E(X_i).$$

Les X_i étant supposés de même loi, on en déduit :

$$\mathbb{E}(N_n) = N_0 \times E(X)^n = N_0 \, \exp[n \times \log \mathbb{E}(X_i)] = N_0 \, e^{r_M n}.$$

Le taux de croissance proposé par Michel correspond à la croissance de la population en moyenne.

5) On suppose que X vaut $0,8$ avec probabilité $0,4$ et $1,1$ avec probabilité $0,6$. Que valent r_M, r_B ? Les conclusions sont-elles identiques quant au risque d'extinction ?

Reponse :$\mathbb{E}(X) = 0,4 \times 0,8 + 0,6 \times 1,1 = 0.98$ donc $r_M = \log(0,98) \approx -2,02.10^{-2} < 0$.

$r_B = \mathbb{E} \log(X) = 0,4 \times \log(0,8) + 0,6 \times \log(1,1) \approx -3,21.10^{-2} < 0$. Même si ces deux définitions ne coïncident pas (avec un facteur $3/2$ entre les deux), elles donnent les même conclusions face au risque d'extinction.

6) Relier $\log(N_n/N_0)$ à une loi binomiale dont on précisera les paramètres.

*Reponse :*Vu la question 1, on peut écrire :

$$\log(N_n/N_0) = \sum_{1}^{n} \log(X_i) = n \, \log(0,8) + \sum_{1}^{n} \log(X_i/0.8)$$

Les variables aléatoires $(\log(X_i/0.8))$ sont indépendantes et valent chacune $\log(0.8/0.8) = 0$ avec probabilité 0,4 et $\log(1.1/0.8)$ avec probabilité 0,6. Si on les divise par $a := \log(1.1/0.8)$, on obtient donc des variables de Bernoulli indépendantes : $Y_i = \log(X_i/0.8)/\log(1.1/0.8)$. Avec $b := \log(0,8)$, on peut donc écrire :

$$\log(N_n/N_0) = n\, b + a \times \sum_{1}^{n} Y_i = n\, b + a\, S_n,$$

où S_n suit une loi binomiale de paramètres n et $p = 0.6$.

Alternative : prendre $\log(1,1)$ comme valeur de b et $a = -\log(1.1/0.8)$ avec S_n une Binomiale de paramètre n et $p = 0,4$ est aussi une bonne réponse.

Remarque : *Toute variable X à valeurs réelle qui ne prend que deux valeurs peut toujours être vue comme de la forme $X = aB + b$ où $a, b \in \mathbb{R}$ et B suit une loi de Bernoulli.*

7) La loi des grands nombres donne-t-elle une justification à une notion de taux de croissance ? Qu'est-ce que cela traduit, notamment vis-à-vis de la question 4 ?

*Reponse :*Par la loi des Grands Nombres, $\sum_1^n \log(X_i)$ évolue lorsque n est grand comme $n\,\mathbb{E}\log(X) = n\,r_B$.

Pour n assez grand, on s'attend donc à retrouver typiquement (i.e. avec une probabilité conséquente) une taille de population de l'ordre de e^{nr_B}. r_B qualifie donc le comportement typiquement observé, qui est différent du comportement moyen car les populations chanceuses croissent très fortement.

Remarque : *Vous pouviez aussi l'obtenir en appliquant le Théorème Central Limite, que ce soit sur $\sum \log(X_i)$ ou sur la loi binomiale si c'est plus facile pour vous.*

8) Rappeler le Théorème de Moivre-Laplace sur les lois binomiales. Peut-on bien l'appliquer à ce cadre ?

*Reponse :*Par le Théorème de Moivre-Laplace (i.e. le Théorème Central Limite dans le cas particulier des binomiales), on a que pour une loi binomiale S_n de paramètre n et p, la variable aléatoire $\frac{S_n - n\,p}{\sqrt{np(1-p)}}$ converge vers une loi normale centrée réduite à la limite où n tend vers l'infini.

L'approximation de cette variable par la loi normale est tout à fait justifiée dès lors que n sera suffisamment grand (de l'ordre de 30 sachant que ni $p = 0,6$, ni $1 - p = 0,4$ ne sont petits).

9) Avec quelle probabilité la taille de population a-t-elle été divisée par plus de 10 au cours des 50 premières générations ?

*Reponse :*On est bien dans le cadre précédent d'application du Théorème de Moivre Laplace, qui nous permet d'approximer $\log(N_n/N_0)$ par

$$n\,b + a\,n\,p + a\,\sqrt{np(1-p)}G = 50 \times r_B + \log(1.1/0.8)\,\sqrt{50 \times 0,6 \times 0,4}G \approx -1.65 + 1.10G.$$

avec G suivant une loi normale centrée réduite. La taille de population a été divisée par plus de 10 si et seulement si $\log(N_n/N_0) \leq -\log(10) = -2.30$. On estime donc la probabilité associée par

$$\mathbb{P}(-1.65 + 1.10\,G \leq -2.30) = \mathbb{P}(G \leq -\frac{2.3 - 1.65}{1.10}) = -\mathbb{P}(G \leq -0.59).$$

La probabilité est tout à fait conséquente (vu que 0.59¡1).

Remarque : Comme vous n'aviez pas les tables de la gaussienne, cela m'aurait clairement suffi. Il se trouve que la probabilité est de l'ordre de 27% !

10) Donner une approximation à densité pour la loi de N_n pour de grandes valeurs de N.

*Reponse :*On écrit $\log(N_n/N_0) \approx n\,r_B + a\sqrt{np(1-p)}G$. On identifie la densité associée en calculant pour une fonction test positive f quelconque :

$$\begin{aligned}\mathbb{E}(f[N_n]) &= \int_{\mathbb{R}} f[N_0 \exp(n\,r_B + a\sqrt{np(1-p)}y)]\frac{1}{\sqrt{2\,\pi}}e^{-y^2/2}\,dy \\ &= \frac{1}{a\sqrt{2\,\pi\,np(1-p)}}\int_{\mathbb{R}_+^*} f[N_0 \exp(z)] \exp\big[\frac{(z - n\,r_B)^2}{2a^2np(1-p)}\big]\,dz, \\ &\qquad \text{où } z := n\,r_B + a\sqrt{np(1-p)}y.\end{aligned}$$

On finit avec le changement de variable $x = N_0 \exp(z)$, qui définit une bijection de R dans R_+^* et correspond à $z = log(x/N_0)$, donc $dz = dx/x$:

$$\mathbb{E}(f[N_n]) = \frac{1}{a\sqrt{2\,\pi\,np(1-p)}}\int_{\mathbb{R}} \frac{f[x]}{x} \exp\big[\frac{(log(x/N_0) - n\,r_B)^2}{2a^2np(1-p)}\big]\,dx.$$

La densité de N_n en x obtenue ainsi est donc estimée par :

$$\frac{1}{ax\sqrt{2\,\pi\,np(1-p)}}\exp\big[\frac{(log(x/N_0)-n\,r_B)^2}{2a^2np(1-p)}\big]$$

pour $x \in \mathbb{R}_+^*$ et 0 sinon.

11) On considère maintenant le cas d'événements catastrophiques, mais rares. On suppose que X vaut 0.01 avec probabilité 0.01 et 1.03 sinon. Que valent r_M, r_B ? Qu'en dîtes-vous ?
Indication : $log(100) \approx 4.61$, $log(1.03) \approx 2.9610^{-2}$.

Reponse :$\mathbb{E}(X) = 0,99 \times 1,03 + 0,01 \times 0,01 \approx 1.0198$ donc $r_M = \log(1.0198) \approx 1.98.10^{-2} > 0$. Cette valeur est pratiquement insensible à ces événements d'extinction massive.

$r_B = \mathbb{E}\log(X) = 0,99 \times \log(1,03) + 0,01 \times \log(0,01) \approx -1,67.10^{-2} < 0$. Les conclusions sont cette fois opposées selon la valeur de taux de croissance choisie : seule Béatrice conclut que l'espèce est en danger.

12) On suppose que N_0 vaut un million et que l'espèce est en grand danger si sa population tombe en-dessous de 100. On admettra le modèle ci-dessus pour l'évolution de N_n tant que ce seuil n'est pas atteint. Tracer quelques réalisations "typiques" de telles dynamiques aléatoires sur 200 générations.

*Reponse :*cf figures en fin de document

13) Evaluer la probabilité que l'espèce devienne en grand danger lors des 200 premières générations.
Indication : $e^{-2} \approx 0.135$.

*Reponse :*Le taux de croissance sans catastrophe (de 0.02) fait croître la population d'au plus $\exp[200 \times 0.02] \approx 50$. S'il y a eu plus de 3 catastrophe qui ont divisé chacune la taille de population par 100, l'espèce est nécessairement atteint un effectif inférieur à 100 lors de la troisième et a amené l'espèce en grand danger. Avec moins de 2 catastrophe, la taille de population est forcément restée supérieure à $10^6/(100)^2 = 100$ et l'espèce n'a pas connu de trop grand risque. Le nombre de catastrophe lors de ces 200 premières générations est donné par une binomiale de paramètres $n = 200$ et $p = 0,01$. Comme p est faible, on peut l'approximer par une loi de Poisson de moyenne $np = 2$. On évalue donc la probabilité demandée à l'aide d'une variable P de poisson de moyenne 2 comme :

$$\mathbb{P}(P \geq 3) = 1 - \mathbb{P}(P = 0) - \mathbb{P}(P = 1) - \mathbb{P}(P = 2) = 1 - e^{-2}(1 + 2 + 4/2) \approx 0,323$$

Il y a presque 1 chance sur 3 que l'espèce ait atteint un niveau de grand danger.

14) Quel est plus généralement l'état de la population après 200 générations ?

*Reponse :*On distingue selon le nombre de catastrophe, puisque l'ordre des X_i n'importe pas (dès lors que l'espèce n'est pas en grand danger). Avec probabilité environ $e^{-2} = 0,135$, aucune catastrophe et l'espèce atteint un effectif de $\exp[n\log(1.03)] \approx 52,5$ millions d'individus.

Avec probabilité environ $2e^{-2} = 0,370$, une seule catastrophe et l'espèce atteint un effectif de $\exp[n\log(1.03)]/100*$ $1000 \approx 525$ milliers d'individus.

Avec probabilité environ $2e^{-2} = 0,370$, deux catastrophes et l'espèce atteint un effectif de $\exp[n\log(1.03)]/(100)2*$ $1000 \approx 5,25$ milliers d'individus.

Lorsque plus de 3 catastrophes ont eu lieu, difficile de prédire exactement mais la plupart des populations ont dû mourir.

15) Que concluez-vous vis-à-vis de ces deux définitions r_M et r_B ?

*Reponse :*Aucune définition n'est pleinement satisfaisante, mais chacune apporte une information intéressante. La mesure de r_B fournit une bonne évaluation de l'évolution du nombre d'individus des populations relativement chanceuses. Mais il néglige assez fortement les événements de déclins, surtout s'ils sont rares (cf les deux questions précédentes). Même pour des événements de déclins fréquents, on voit en question 9 que cette mesure de croissance ne correspond pas à l'évolution de la population telle que fréquemment observée sur de nombreuses générations.

A ce titre, r_B semble un meilleur prédicteur du destin à long terme de la population. Néanmoins, on voit qu'il et très sensible à l'intensité d'événements rares mais catastrophiques. C'est raisonnable sur des échelles de temps où on s'attend à les voir advenir. Si la probabilité d'avoir une telle catastrophe est très faible sur cette période de temps d'intérêt, il vaut mieux soit négliger ces événements potentiels, soit distinguer les cas selon l'arrivée ou non de tels événements (nombre de catastrophes et leurs instants à inclure comme paramètres).

Remarque : Une version assez similaire de ces résultats est parue dans un Pour la Science récent, dans un article consacré à la concentration de richesse. Je vous invite à trouver ce paradoxe du "vide-grenier" dans l'article consacré du numéro 507.

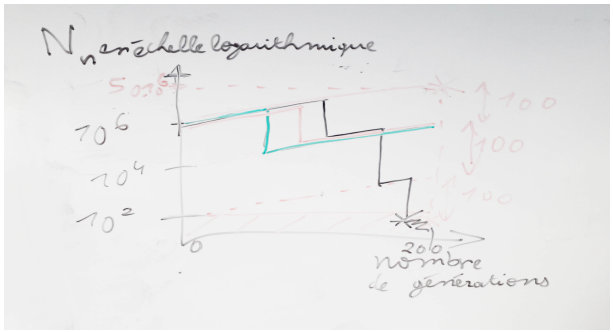

Figure 1: en échelle log pour mieux interpréter

Et quelques images pour la dynamique en échelle "naturelle" :