Corrigé du Devoir

Exercice 1 : jeu de dé

On considère le jeu suivant : on lance un dé ; si on obtient un 5 ou un 6, on gagne 2 et sinon on perd 1. Le jeu peut être répété autant de fois que l'on veut et les gains s'accumulent.

On souhaite comparer deux stratégies de jeu. La première stratégie consiste simplement à lancer 3 fois le dé. La seconde stratégie consiste à lancer le dé jusqu'à ce qu'on obtienne un 5 ou un 6 puis à s'arrêter de jouer.

On note G_1 (respectivement G_2) la variable aléatoire représentant le gain obtenu en suivant la stratégie 1 (respectivement 2).

Déterminer et représenter les densités des variables G_1 et G_2 puis calculer leur espérance et leur variance.

Conclure à partir de tous ces éléments : quelle est la stratégie la plus intéressante pour le joueur ?

On donne les sommes suivantes :

$$\sum_{k=1}^{+\infty} k\left(\frac{2}{3}\right)^{k-1} = 9, \qquad \sum_{k=1}^{+\infty} k^2\left(\frac{2}{3}\right)^{k-1} = 45.$$

Pour toute l'étude, nous supposerons les dés non pipés et les différents lancers indépendants.

Étudions la variable G_1. Notons Y_1 le nombre de parties gagnées sur les quatre parties jouées. La loi de Y_1 est la loi binomiale $\mathcal{B}(3, \frac{1}{3})$ où $\frac{1}{3}$ représente la probabilité qu'un lancer soit victorieux. Le gain est de 3 pour chaque partie gagnée et de -1 pour chaque partie perdue, donc le gain total est donné par $G_1 = 2Y_1 - (3 - Y_1) = 3Y_1 - 3$. On en déduit la densité de G_1 : les valeurs prises par G_1 sont $-3, 0, 3$ et 6 et leurs probabilités sont

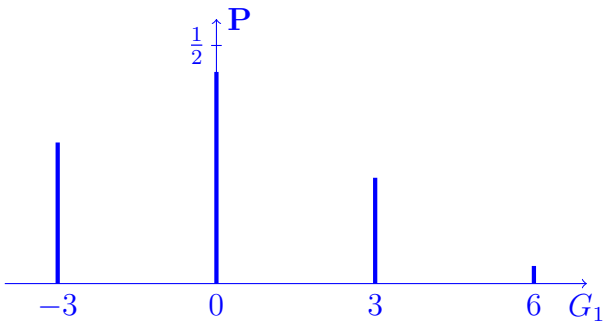
$$\mathbf{P}(G_1 = -3) = \left(\frac{2}{3}\right)^3 = \frac{8}{27}, \quad \mathbf{P}(G_1 = 0) = 3\left(\frac{2}{3}\right)^2\frac{1}{3} = \frac{12}{27},$$
$$\mathbf{P}(G_1 = 3) = 3\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^2 = \frac{6}{27}, \quad \mathbf{P}(G_1 = 3) = \left(\frac{1}{3}\right)^3 = \frac{1}{27}.$$

Espérance de G_1 :

$$\mathbf{E}(G_1) = -3\mathbf{P}(-3) + 0\mathbf{P}(0) + 3\mathbf{P}(3) + 6\mathbf{P}(6) = \frac{1}{27}(-24 + 0 + 18 + 6) = 0.$$

Le jeu est donc équilibré.
Variance de G_1 :

$$\mathbf{V}(G_1) = (-3)^2\mathbf{P}(-3) + 0^2\mathbf{P}(0) + 3^2\mathbf{P}(3) + 6^2\mathbf{P}(6) = \frac{1}{27}(72 + 0 + 54 + 36) = 6.$$



Passons à l'étude de G_2. Notons Y_2 le nombre de lancers effectués jusqu'à l'obtention d'un 5 ou d'un 6. La loi de Y_2 est la loi géométrique $\mathcal{G}(\frac{1}{3})$. Si $Y_2 = k$ alors le gain total est $2-(k-1)$ ($(k-1)$ parties perdues puis une partie gagnée). Ainsi les valeurs possibles pour G_2 sont tous les entiers inférieurs à 2 et

$$\forall k \geqslant 1, \quad \mathbf{P}(G_2 = 3-k) = \mathbf{P}(Y_2 = k) = \left(\frac{2}{3}\right)^{k-1} \frac{1}{3},$$

ou encore

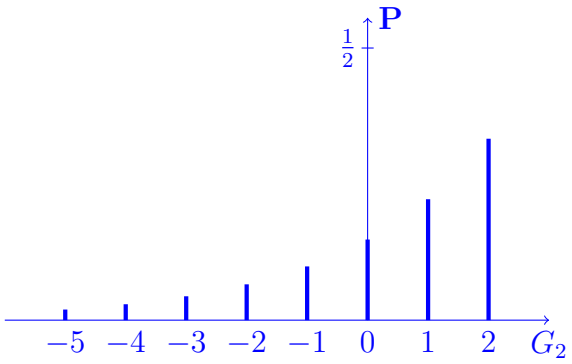
$$\forall a \leqslant 2, \quad \mathbf{P}(G_2 = a) = \mathbf{P}(Y_2 = 3-a) = \left(\frac{2}{3}\right)^{2-a} \frac{1}{3}.$$

Espérance de G_2 :

$$\mathbf{E}(G_2) = \sum_{k=1}^{+\infty} (3-k) \left(\frac{2}{3}\right)^{k-1} \frac{1}{3} = 3 - \frac{9}{3} = 0.$$

Variance de G_2 :

$$\mathbf{V}(G_2) = \sum_{k=1}^{+\infty} (3-k)^2 \left(\frac{2}{3}\right)^{k-1} \frac{1}{3} = 9 - 6\frac{9}{3} + \frac{45}{3} = 6.$$



Conclusion : avec ces deux stratégies, le gain moyen est le même avec des variances également identiques. On ne peut donc pas dire que l'une est plus avantageuse que l'autre. Si ce jeu était répété un grand nombre de fois, les gains totaux seraient sensiblement identiques.

On peut tout de même préciser les différences de ces deux stratégies en étudiant leurs densités. Avec la première stratégie, on est gagnant avec une probabilité $\frac{7}{27}$ tandis qu'avec la seconde, cette probabilité est de $\frac{1}{3} + \frac{2}{9} = \frac{15}{27}$. On a donc plus de chances de gagner avec la stratégie 2. En revanche, on peut espérer gagner jusqu'à 6 avec la stratégie 1 tandis que le gain maximal est de 2 avec la seconde. Enfin, la perte maximale est de -3 avec la stratégie 1 et n'est pas limitée avec la stratégie 2.

Finalement, on peut dire que sur le long terme, les deux stratégies sont équivalentes, mais sur le court terme, la stratégie 1 a moins de chances d'être gagnante que la seconde mais fait prendre moins de risques et peut faire gagner plus que la seconde. Dire laquelle des deux est la plus intéressante n'est qu'affaire de choix.

Exercice 2 : trajectoire aléatoire

On considère un point se déplaçant aléatoirement sur un intervalle de la manière suivante : chaque nouvelle position du point est choisie aléatoirement dans l'intervalle, indépendamment des positions précédentes. On souhaite décrire la distance parcourue par le point au bout d'un certain nombre d'étapes.

1. **Distance aléatoire**

   Soient X_1 et X_2 deux variables indépendantes de même loi uniforme $\mathcal{U}([0,1])$. On pose $Y = |X_2 - X_1|$.

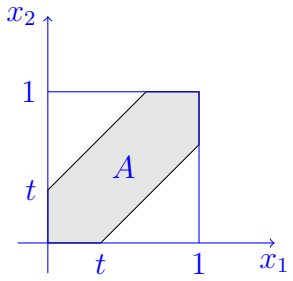
   (a) Déterminer la fonction de répartition de Y et en déduire que sa densité est donnée par la fonction

   $$f_Y(x) = 2 - 2x \qquad \text{pour } x \in [0,1].$$

   Les variables X_1 et X_2 étant à valeurs dans $[0,1]$, Y est également à valeurs dans $[0,1]$. Soit $t \in [0,1]$. Alors

   $$F_Y(t) = \mathbf{P}(Y \leqslant t) = \mathbf{P}(-t \leqslant X_2 - X_1 \leqslant t).$$

   Représentons l'ensemble A des couples $(x_1, x_2) \in [0,1]^2$ tels que $-t \leqslant x_2 - x_1 \leqslant t$.

   Comme X_1 et X_2 sont des variables indépendantes de densité 1 sur $[0,1]$, la mesure de A est

   $$\mathbf{P}(A) = \iint_A 1 \times 1 dx_1 dx_2 = \mathcal{A}ire(A).$$

   Ainsi $F_Y(t) = 1 - 2\frac{(1-t)^2}{2} = 1 - (1-t)^2$.

   Et la densité de Y est donnée pour $t \in [0,1]$ par

   $$f_Y(t) = F_Y'(t) = 2(1-t).$$

   (b) Calculer l'espérance et la variance de Y. *On doit trouver un écart-type* $\sigma_Y \approx 0,24$.

   $$\mathbf{E}(Y) = \int_0^1 t \times 2(1-t)dt = \frac{1}{3}, \qquad V(Y) = \int_0^1 (t - \frac{1}{3})^2 \times 2(1-t)dt = \frac{1}{18}.$$

2. **Distance totale**

   On considère maintenant la trajectoire du point. Soient $(X_i)_i \in \mathbf{N}^*$ une suite de variables indépendantes de loi $\mathcal{U}([0,1])$: ce sont les positions du point au cours du mouvement. Pour $i \geqslant 1$, notons $Y_i = |X_{i+1} - X_i|$. Notons enfin $S_n = \sum_{i=1}^n Y_i$ la distance parcourue par le point au bout de n étapes.

(a) Pourquoi ne peut-on pas appliquer le théorème central limite aux variables Y_i ?

Les variables Y_1 et Y_2 sont définies par $Y_1 = |X_2 - X_1|$ et $Y_2 = |X_3 - X_2|$. La variable X_2 intervenant dans les deux expressions, Y_1 et Y_2 sont très certainement non indépendantes. (On peut par exemple montrer que $\mathbf{P}(Y_2 \geqslant \frac{1}{2} \mid Y_1 = 1) = \frac{1}{2} \neq \mathbf{P}(Y_2 \geqslant \frac{1}{2}) = \frac{1}{4}$.)

Nous allons tout de même appliquer illicitement ce théorème aux variables Y_i et nous verrons dans la partie suivante comment le faire rigoureusement.

(b) Écrire le théorème central limite en faisant apparaître la variable S_n.

En faisant comme si les variables Y_i étaient indépendantes, on pourrait écrire pour $n \geqslant 30$ que $\frac{S_n - \frac{n}{3}}{\sigma\sqrt{n}}$ (avec $\sigma = \frac{1}{\sqrt{18}} \approx 0,24$) suit à peu près la loi $\mathcal{N}(0,1)$.

(c) Pour $n = 60$, déterminer approximativement $\mathbf{P}(S_n \geqslant 17)$.

Pour $n = 60$, on a $\frac{S_n - 20}{1,83} \sim \mathcal{N}(0,1)$. Alors

$$\mathbf{P}(S_n \geqslant 17) = \mathbf{P}(\frac{S_n - 20}{1,83} \geqslant -1,64) \approx 1 - F_{\mathcal{N}}(-1,64) = F_{\mathcal{N}}(1,64) \approx 0,95.$$

(d) Pour $n = 120$, déterminer approximativement $\mathbf{P}(S_n \geqslant 34)$.

Cette fois $\frac{S_n - 40}{2,58} \sim \mathcal{N}(0,1)$ et

$$\mathbf{P}(S_n \geqslant 34) = \mathbf{P}(\frac{S_n - 40}{2,58} \geqslant -2,32) \approx F_{\mathcal{N}}(2,32) \approx 0,99.$$

(e) Pour $n = 150$, déterminer a tel que $\mathbf{P}(|S_n - 50| \leqslant a) \approx 0,95$.

Cette fois $\frac{S_n - 50}{2,89} \sim \mathcal{N}(0,1)$ et

$$\mathbf{P}(|S_n - 50| \leqslant a) = \mathbf{P}(-\frac{a}{2,89} \leqslant \frac{S_n - 50}{2,89} \leqslant \frac{a}{2,89}) \approx 2F_{\mathcal{N}}(\frac{a}{2,89}) - 1.$$

On veut que cette probabilité soit de l'ordre de $0,95$. Il faut donc que $F_{\mathcal{N}}(\frac{a}{2,89}) \approx 1,96$. On en déduit $a \approx 5,7$.
Ainsi au risque de 5%, S_n est compris entre $44,3$ et $55,7$.

(f) Déterminer la valeur de n à partir de laquelle on peut affirmer que la distance parcourue est supérieure à 100 avec un risque d'erreur de 1%.

On cherche n tel que $\mathbf{P}(S_n \geqslant 100) \approx 0,99$. Or

$$\mathbf{P}(S_n \geqslant 100) = \mathbf{P}(\frac{S_n - \frac{n}{3}}{\sigma\sqrt{n}} \geqslant \frac{100 - \frac{n}{3}}{\sigma\sqrt{n}}) \approx 1 - F_{\mathcal{N}}(\frac{100 - \frac{n}{3}}{\sigma\sqrt{n}}).$$

Pour que cette probabilité soit égale à 0,99, il faut avoir $\frac{100 - \frac{n}{3}}{\sigma\sqrt{n}} \approx -2,33$, donc $\frac{n}{3} - 2,33\sigma\sqrt{n} - 100 = 0$. On résout cette équation de degré 2 en $\sqrt{n}$ et on trouve $n \approx 330$.

3. **Estimations rigoureuses**

Les résultats obtenus dans la partie précédente sont faux à cause de l'utilisation erronée du théorème central limite. Nous allons ici établir des résultats plus rigoureux en s'appuyant toujours sur ce théorème.

(a) Soient A et B des événements aléatoires. Montrer que $\mathbf{P}(A\cap B) \geqslant \mathbf{P}(A)+\mathbf{P}(B)-1$.

On sait que $\mathbf{P}(A \cup B) = \mathbf{P}(a) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. Comme une probabilité est toujours inférieure à 1, $\mathbf{P}(a) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \leqslant 1$ et le résultat en découle.

(b) Séparer la somme S_n en deux sous-sommes de variables Y_i pour lesquelles le théorème central limite est applicable.

On a vu que les variables Y_i ne sont pas indépendantes. Cependant certaines le sont. Par exemple, comme les variables X_i sont indépendantes, $Y_1 = |X_2 - X_1|$ et $Y_3 = |X_4 - X_3|$ sont indépendantes. Et de manière générales toutes les variables Y_i pour i impair sont deux à deux indépendantes, et toutes les variables Y_i pour i pair sont deux à deux indépendantes. Ainsi si on écrit $S_n = \sum_{i \text{ impair } \leqslant n} Y_i + \sum_{i \text{ pair } \leqslant n} Y_i = S_n^{(i)} + S_n^{(p)}$, on reconnaît des sommes de variables aléatoires indépendantes de même loi et le TCL va pouvoir s'appliquer à chacune de ces sommes.

(c) En utilisant un des calculs de la partie précédente, estimer de nouveau la probabilité $\mathbf{P}(S_n \geqslant 34)$ pour $n = 120$ et commenter.

Si $S_n^{(i)}$ et $S_n^{(p)}$ sont supérieures à 17, alors la somme S_n est supérieure à 34 (la réciproque est fausse). On en déduit que $\mathbf{P}(S_n \geqslant 34) \geqslant \mathbf{P}(S_n^{(i)} \geqslant 17 \text{ et } S_n^{(p)} \geqslant 17)$. Et d'après la question précédente

$$\mathbf{P}(S_n \geqslant 34) \geqslant \mathbf{P}(S_n^{(i)} \geqslant 17) + \mathbf{P}(S_n^{(p)} \geqslant 17) - 1.$$

En utilisant le TCL, on a déjà estimé la probabilité qu'une somme de 60 variables Y_i indépendantes soit supérieure à 17. Ce calcul est parfaitement rigoureux si on raisonne avec des variables indépendantes. Comme c'est le cas des deux sommes $S_n^{(i)}$ et $S_n^{(p)}$, on peut affirmer rigoureusement que $\mathbf{P}(S_n^{(i)} \geqslant 17) \approx 0,95$ et $\mathbf{P}(S_n^{(p)} \geqslant 17) \approx 0,95$. on en déduit que pour $n = 120$

$$\mathbf{P}(S_n \geqslant 34) \geqslant 0,90.$$

(d) Déterminer de même la valeur de n à partir de laquelle on peut affirmer que la distance parcourue est supérieure à 200 avec un risque d'erreur d'au plus 2%.

Raisonnons de la même façon et utilisons le résultat de la question $2-f$.

$$\mathbf{P}(S_n \geqslant 200) \geqslant \mathbf{P}(S_n^{(i)} \geqslant 100 \text{ et } S_n^{(p)} \geqslant 100) \geqslant \mathbf{P}(S_n^{(i)} \geqslant 100) + \mathbf{P}(S_n^{(p)} \geqslant 100) - 1.$$

Or pour des variables indépendantes, on a démontré qu'à partir de 330 termes, une somme de variables Y_i indépendantes était supérieure à 100 au risque de 1%. Donc pour $n = 660$, $\mathbf{P}(S_n^{(i)} \geqslant 100) = \mathbf{P}(S_n^{(p)} \geqslant 100) \approx 0,99$. Et finalement on peut affirmer que pour $n \geqslant 660$,

$$\mathbf{P}(S_n \geqslant 200) \geqslant 0,98.$$

Remarque : dans cette dernière partie, on obtient des minorations des probabilités recherchées, pas des estimations précises. Ces résultats ont le mérite d'être justes alors que ceux obtenus dans la seconde partie étaient faux.

Pour obtenir des estimations précises, il faudrait étudier en détail la dépendance entre les variables Y_i paires et impaires.

On peut également les estimer à partir de simulations numériques. On a obtenu de cette façon les résultats suivants :

$$\mathbf{P}(S_{60} \geqslant 17) \approx 0,93 - 0,94 \qquad \mathbf{P}(S_{120} \geqslant 34) \approx 0,98 - 0,99 \qquad \mathbf{P}(S_{330} \geqslant 100) \approx 0,98 - 0,99.$$