Exercice non corrigé en cours

Exercice 1 – *Problème du collectionneur de vignettes.* Chaque paquet de céréales contient une vignette à collectionner, que l'on ne découvre qu'à l'ouverture du paquet. On se demande combien il faut ouvrir de paquets pour posséder au moins un exemplaire de chacune des n vignettes.
On décompose ce nombre en $N_n = \tau_1 + \tau_2 + \cdots + \tau_n$ où τ_k est le nombre de paquets supplémentaires nécessaires pour obtenir k vignettes différentes quand on en a déjà $k-1$ différentes.

1. Justifier que τ_k suit une loi géométrique, dont on précisera le paramètre.

À partir du moment où $k-1$ vignettes différentes ont été obtenues, chaque ouverture de paquet a une probabilité $\frac{n-(k-1)}{n}$ de fournir une nouvelle vignette (en supposant chaque vignette équiprobable), et les vignettes sont (implicitement) supposées indépendantes d'un paquet à l'autre, donc (par la proposition du cours qui suit la définition de la loi géométrique) le nombre de paquets à ouvrir pour avoir une vignette différente suit une loi géométrique de paramètre $\frac{n-(k-1)}{n}$: c'est l'instant du premier « succès » dans une suite d'épreuves indépendantes où un succès a pour probabilité $\frac{n-(k-1)}{n}$.

2. En déduire l'espérance de N_n (en donner une valeur approchée).

On a donc $E[\tau_k] = \frac{n}{n-k+1}$, et ainsi par linéarité de l'espérance

$$E[N_n] = E[\tau_1] + \cdots + E[\tau_n] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1}\right) = n\sum_{k=1}^{n}\frac{1}{k}$$

(pour la dernière égalité, on a juste inversé l'ordre des termes).
Or on sait que (par comparaison série-intégrale) $\sum_{k=1}^{n} \frac{1}{k} \sim_n \ln n$, d'où

$$E[N_n] \sim_n n \ln n.$$

3. On admet que les variables aléatoires $(\tau_k)_{1\leq k\leq n}$ sont indépendantes (le justifier intuitivement).

Pour $k = 1, \ldots, n$, la valeur de τ_k ne dépend que
– des vignettes obtenues après le moment où on a $k-1$ vignettes différentes, et
– des $k-1$ types de vignettes déjà obtenus (afin de savoir si une vignette est nouvelle).
Or, d'une part, les tirages sont indépendants, donc les vignettes obtenues après en avoir eu $k-1$ différentes sont indépendantes des précédentes, et donc de $\tau_1, \ldots, \tau_{k-1}$. Et, d'autre part, connaître les $k-1$ premiers types de vignettes obtenus ne renseigne pas sur les temps $\tau_1, \ldots, \tau_{k-1}$ mis à les obtenir, puisque chaque type de vignette est équiprobable. Par suite, τ_k est indépendant de $\tau_1, \ldots, \tau_{k-1}$. Par récurrence, on en déduit que $\tau_1, \ldots, \tau_n$ sont indépendantes.

3.a) En déduire la variance de N_n, et montrer que $\mathrm{Var}(N_n) \leq Cn^2$ pour une constante C.

Comme $\tau_1, \ldots, \tau_n$ sont indépendantes,

$$\mathrm{Var}(N_n) = \mathrm{Var}(\tau_1) + \cdots + \mathrm{Var}(\tau_n).$$

Or la variance d'une variable de loi géométrique de paramètre p est $\frac{1-p}{p^2}$, d'où

$$\begin{aligned}
\mathrm{Var}(N_n) &= \frac{1-\frac{n}{n}}{\left(\frac{n}{n}\right)^2} + \frac{1-\frac{n-1}{n}}{\left(\frac{n-1}{n}\right)^2} + \cdots + \frac{1-\frac{2}{n}}{\left(\frac{2}{n}\right)^2} + \frac{1-\frac{1}{n}}{\left(\frac{1}{n}\right)^2} \\
&= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + \cdots + \frac{n-2}{2^2} + \frac{n-1}{1^2}\right) \\
&\leq n^2\left(\frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} + \cdots + \frac{1}{2^2} + \frac{1}{1^2}\right) \\
&\leq n^2 \sum_{k=1}^{\infty} \frac{1}{k^2}
\end{aligned}$$

et la dernière série converge, donc on a obtenu la majoration voulue.

3.b) En déduire, pour tout $\varepsilon > 0$,

$$P\Big(\big|N_n - E[N_n]\big| > \varepsilon n \log n\Big) \underset{n}{\longrightarrow} 0,$$

puis

$$\frac{N_n}{n \ln n} \xrightarrow[n]{(p)} 1.$$

Appliquons l'inégalité de Tchebychev à N_n : pour tout $\varepsilon > 0$,

$$\begin{aligned} P\Big(\big|N_n - E[N_n]\big| > \varepsilon n \ln n\Big) &\leq \frac{\mathrm{Var}(N_n)}{\varepsilon^2 n^2 (\ln n)^2} \\ &\leq \frac{C}{\varepsilon^2 (\ln n)^2} \underset{n}{\longrightarrow} 0. \end{aligned} \qquad (*)$$

C'est le premier point. Pour obtenir la convergence en probabilité demandée, il faut maintenant montrer que, pour tout $\varepsilon > 0$, $P(|\frac{N_n}{n \ln n} - 1| > \varepsilon) \to 0$.
Soit $\varepsilon > 0$. On rappelle qu'on a vu que $\frac{E[N_n]}{\ln n} \to 1$, donc il existe $n_0 > 0$ tel que, si $n \geq n_0$, $\left|\frac{E[N_n]}{n \ln n} - 1\right| < \frac{\epsilon}{2}$ et donc, pour $n \geq n_0$,

$$\begin{aligned} \left|\frac{N_n}{n \ln n} - 1\right| &\leq \left|\frac{N_n}{n \ln n} - \frac{E[N_n]}{n \ln n}\right| + \left|\frac{E[N_n]}{n \ln n} - 1\right| \\ &\leq \frac{\big|N_n - E[N_n]\big|}{n \ln n} + \frac{\varepsilon}{2}. \end{aligned}$$

Ainsi, pour $n \geq n_0$, si $\left|\frac{N_n}{n \ln n} - 1\right| > \varepsilon$, alors $\frac{|N_n - E[N_n]|}{n \ln n} > \varepsilon - \frac{\varepsilon}{2} = \frac{\varepsilon}{2}$. Donc, pour $n \geq n_0$,

$$P\left(\left|\frac{N_n}{n \ln n} - 1\right| > \varepsilon\right) \leq P\left(\big|N_n - E[N_n]\big| \geq \frac{\varepsilon}{2} n \ln n\right)$$

(l'événement de gauche implique celui de droite, donc est inclus dans celui-ci, d'où l'inégalité) et cette dernière quantité converge vers 0 quand $n \to \infty$ par $(*)$.