Travaux dirigés, feuille 4

Corrigé succint (par Francis Comets)

Partie A:

1. On peut introduire la famille de variables $(Y_i)_{i\geq 1}$ où pour tout $i \geq 1$, $Y_i \in \{0, ..., i-1\}$: $Y_i = k$, $k \in 1, ..., i-1$ si le $i$ème individu choisit la table de l'individu k, et $Y_i = 0$ si le $i$ème individu choisit une nouvelle table. D'après l'énoncé, on sait que la famille $(Y_i)_{i\geq 1}$ est indépendante, que $\mathbf{P}(Y_i = k) = \frac{1}{i-1+\theta}$ si $k \in \{1, ..., i-1\}$ et $\mathbf{P}(Y_i = 0) = \frac{\theta}{i-1+\theta}$. On a alors

$$p_{n+1,1} = \mathbf{P}[K_{n+1} = 1] = \mathbf{P}[K_n = 1, Y_{n+1} = 0] = \mathbf{P}[K_n = 1]\mathbf{P}[Y_{n+1} = 0] = p_{n,1} \times \frac{n}{n+\theta}\,,$$

où on a utilisé le fait que K_n, étant $\sigma(Y_1, \ldots, Y_n)$-mesurable, est indépendant de Y_{n+1}. On obtient donc

$$p_{n+1,1} = \frac{1}{1+\theta} \times \frac{2}{2+\theta} \times \ldots \frac{n}{n+\theta}\,.$$

De même,

$$\begin{aligned} p_{n+1,i} &= \mathbf{P}(K_{n+1} = i) \\ &= \mathbf{P}(K_n = i, Y_{n+1} \neq 0) + \mathbf{P}(K_n = i-1, Y_{n+1} = 0) \\ &= \frac{n}{n+\theta}\, p_{n,i} + \frac{\theta}{n+\theta}\, p_{n,i-1} \end{aligned}$$

2. On a

$$\begin{aligned} P_{n+1}(x) &= \sum_{i=1}^{n+1} p_{n+1,i}\, x^i \\ &= p_{n+1,1}\, x + \sum_{i=2}^{n} p_{n+1,i}\, x^i + p_{n+1,n+1}\, x^{n+1}\,. \end{aligned}$$

On applique alors la formule de récurrence de la question 1), et on utilise l'égalité $p_{n+1,n+1} = \theta^n/((1+\theta)\ldots(n+\theta))$, pour obtenir

$$P_{n+1}(x) = \frac{n+\theta x}{n+\theta}\, P_n(x)\,.$$

Puisque $P_1(x) = x$, on en déduit que $P_n(x) = R_n(\theta x)/R_n(\theta)$.

3. Puisque P_n est la fonction génératrice de K_n et que $P_n(1) = 1$, on a

$$\mathbf{E}K_n = P_n'(1) = \frac{P_n'}{P_n}(1) = (\ln P_n)'(1) = \big(\sum_{i=0}^{n-1} \ln(\theta x + i)\big)'(1) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \tag{1}$$

Semblablement, $\mathrm{Var}(K_n) = P_n''(1) + P_n'(1) - P_n'(1)^2$, et en utilisant les dérivées logarithmiques,

$$P_n'(x) = \Big(\sum_{i=0}^{n-1} \frac{\theta}{\theta x + i}\Big) P_n(x) \ ,$$

donc

$$P_n''(1) = -\sum_{i=0}^{n-1} \Big(\frac{\theta}{\theta + i}\Big)^2 + \Big(\sum_{i=0}^{n-1} \frac{\theta}{\theta + i}\Big)^2 \ ,$$

et finalement

$$\mathrm{Var}(K_n) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i}\Big(1 - \frac{\theta}{\theta + i}\Big) = \sum_{i=1}^{n-1} \frac{i\theta}{(\theta + i)^2} \tag{2}$$

Partie B:

1. Au convive numéro i on associe une variable X_i valant 1 ou 0 selon que i choisit une nouvelle table ou non, cad que $X_i = \mathbf{1}_{Y_i = 0}$. Ainsi, X_i est une variable de Bernoulli de paramètre

$$q_i := \frac{\theta}{\theta + i - 1} \ .$$

On a $K_n = X_1 + \ldots + X_n$, et les X_i sont indépendantes par hypothèse.

2. On a $\mathbf{E}K_n = \sum_{i=1}^n \mathbf{E}X_i$, et $\mathrm{Var}(K_n) = \sum_{i=1}^n \mathrm{Var}(X_i)$ par indépendance. Avec $\mathbf{E}X_i = q_i$ et $\mathrm{Var}(X_i) = q_i(1 - q_i)$, on retrouve les résultats de (1) et (2).

3. Par monotonie de $x \mapsto \theta/(\theta + x)$, on a

$$\frac{\theta}{\theta + i} \le \int_{i-1}^{i} \frac{\theta}{\theta + x} dx \le \frac{\theta}{\theta + i - 1} \ ,$$

ce qui donne l'encadrement voulu. Puisque

$$\int_0^n \frac{\theta}{\theta + x} dx = \theta\big[\ln(\theta + x)\big]_0^n = \theta \ln(1 + n/\theta) \sim \theta \ln n$$

quand $n \to \infty$, on en déduit que

$$\mathbf{E}K_n \sim \theta \ln n, \quad n \to \infty.$$

4. Puisque $\mathrm{Var}(X_i) = q_i(1-q_i) \le q_i = \mathbf{E}X_i$, on a $\mathrm{Var}(K_n) \le \mathbf{E}K_n$. Ainsi,

$$\begin{aligned} \|\frac{K_n}{\ln n} - \theta\|_2^2 &= \mathbf{E}\Big(\frac{K_n}{\ln n} - \theta\Big)^2 \\ &= \frac{\mathrm{Var}(K_n)}{\ln^2 n} + \Big(\frac{\mathbf{E}K_n}{\ln n} - \theta\Big)^2 \\ &= \mathcal{O}(1/\ln n) + o(1) \end{aligned}$$

tend vers 0 quand n tend vers l'infini. Par conséquent, $K_n/\ln n \to \theta$ dans L^2 et donc aussi en probabilité.

5. Par indépendance des sommants dans la formule $K_n = \sum_{i=1}^n X_i$ et puisque les X_i sont de loi de Bernoulli, la fonction caractéristique de K_n est donnée par

$$\Phi_{K_n}(t) = \prod_{k=1}^n \Phi_{X_k}(t) = \prod_{k=1}^n \big[1 + q_k(e^{it} - 1)\big] \ .$$

Puisque $c_n \to \infty$, on a les développements limités pour tout $t \in \mathbb{R}$,

$$1 + q_k(e^{it/c_n} - 1) = 1 + q_k\left(i\frac{t}{c_n} - \frac{t^2}{2c_n^2} + \mathcal{O}(c_n^{-3})\right) \ , \tag{3}$$

$$\log\big[1 + q_k(e^{it/c_n} - 1)\big] = q_k\left(i\frac{t}{c_n} - \frac{t^2}{2c_n^2}(1-q_k) + \mathcal{O}(c_n^{-3})\right) \ , \tag{4}$$

$$\log \Phi_{K_n - \mathbf{E}K_n}(t/c_n) = -\frac{t^2}{2c_n^2}\sum_{k=1}^n q_k(1-q_k) + \mathcal{O}(c_n^{-3})\sum_{k=1}^n q_k \ . \tag{5}$$

On notera que dans les équations (3), (4) et (5), les $\mathcal{O}(c_n^{-3})$ sont uniformes en k : c'est évident dans (3); ça reste vrai dans (4) car on a la majoration $|q_k| \le 1$; on peut donc mettre en facteur ce $\mathcal{O}(c_n^{-3})$, uniforme en k, devant la somme des q_k dans (5). On est conduit à poser

$$c_n = \sqrt{\sum_{i=1}^n q_i(1-q_i)} = \sqrt{\sum_{i=1}^{n-1} \frac{\theta i}{(\theta+i)^2}} \ .$$

Puisque $\sum_{i=1}^n q_i \sim \theta \ln n$ et $\sum_{i\ge 1} q_i^2 < \infty$, cette suite c_n convient. De plus, on a encore

$$c_n \sim \sqrt{\theta \ln n} \ , \tag{6}$$

quand $n \to \infty$. D'après (5) et (6), il vient

$$\lim_{n\to\infty} \Phi_{\frac{K_n - \mathbf{E}K_n}{\sqrt{\log n}}}(t) = \exp(-\theta t^2/2) \ .$$

Comme la convergence des fonctions caractéristiques entraîne la convergence en loi des variables aléatoires, on obtient que

$$\frac{K_n - \mathbf{E}K_n}{\sqrt{\log n}} \longrightarrow Z \qquad \text{en loi,}$$

avec Z une variable gaussienne $\mathcal{N}(0,\theta)$.

Commentaire: Ce modèle n'est pas seulement la description de l'occupation des tables à la cantine. En écologie et ou génétique, on s'interesse au nombre et à l'abondance des espèces; la variable K_n apparait naturellement pour le nombre d'espèces dans un échantillon de n individus.