Statistique Mathématique

Corrigé de l'examen — 5 mai 2021

I.1. On considère $B_n = B$ fixé. L'énoncé précise seulement que le citoyen rejoint un parti existant avec probabilité proportionnelle à sa taille et fonde son propre parti avec probabilité proportionnelle à θ. Il s'agit donc simplement de trouver la constante de proportionnalité, disons Z_n. On utilise le fait que les probabilités des différents événements doivent sommer à 1 :

$$\begin{aligned} 1 &= \sum_{b\in B} \frac{|b|}{Z_n} + \frac{\theta}{Z_n} \\ &= \frac{1}{Z_n}\left(\sum_{b\in B} |b| + \theta\right) \\ &= \frac{n+\theta}{Z_n}\,, \end{aligned}$$

où l'on a utilisé le fait que B est une partition de $\{1,\dots,n\}$.

I.2. Si $\theta = 0$, la probabilité d'apparition d'un nouveau parti politique est nulle. La suite $(B_n)_{n\geq 1}$ est donc stationnaire : $B_n = \{\{1\}\}$ pour tout $n \geq 1$. C'est le régime *parti unique*. Au contraire, si $\theta \to +\infty$, alors d'après la question I.1., la probabilité de création d'un nouveau parti politique est 1. On observe donc $B_n = \{\{1\},\dots,\{n\}\}$ pour tout $n \geq 1$, soit n *micro-partis*.

I.3. Les principales limitations du modèle est l'impossibilité pour un citoyen de quitter un parti une fois qu'il l'a rejoint. Il n'y a pas non plus de mécanisme réaliste de durée de vie des citoyens dans ce modèle.

II.1. On définit la variable aléatoire ξ_k comme l'indicatrice de l'événement "le citoyen k crée son propre parti." Autrement dit, $\xi_k = \mathbb{1}_{\{k\}\in B_k}$. D'après la question I.1., chaque ξ_k est une Bernoulli de paramètre $\frac{\theta}{k-1+\theta}$. Comme le cardinal de B_n augmente de 1 si, et seulement si, le citoyen $n+1$ crée son propre parti, on a bien que

$$|B_n| = \xi_1 + \cdots + \xi_n\,.$$

Toujours d'après la question I.1., ce sont des variables aléatoires indépendantes.

II.2. Soit $n \geq 1$. D'après la question II.1.,

$$\begin{aligned} \mathbb{E}\left[|B_n|\right] &= \mathbb{E}\left[\xi_1 + \cdots + \xi_n\right] && (1)\\ &= \sum_{k=1}^{n} \mathbb{E}\left[\xi_k\right] && (2)\\ &= \sum_{k=1}^{n} \frac{\theta}{k-1+\theta}\,. && (3) \end{aligned}$$

On obtient le premier résultat par un simple changement d'indices. Pour la variance, on écrit :

$$\begin{aligned} \mathrm{Var}\left(|B_n|\right) &= \mathrm{Var}\left(\sum_{k=1}^{n} \xi_k\right) \\ &= \sum_{k=1}^{n} \mathrm{Var}\left(\xi_k\right) && \text{(indépendance)} \\ &= \sum_{k=1}^{n} \frac{\theta}{k-1+\theta} \cdot \left(1 - \frac{\theta}{k-1+\theta}\right) \\ &= \sum_{k=1}^{n} \frac{(k-1)\theta}{(k-1+\theta)^2}\,. \end{aligned}$$

On conclut de même.

II.3. L'expression obtenue pour $\mathbb{E}\left[|B_n|\right]$ fait penser à une série harmonique, on se lance dans une comparaison série-intégrale. On écrit tout d'abord

$$\int_k^{k+1} \frac{\mathrm{d}x}{x+\theta} \leq \frac{1}{k+\theta} \leq \int_{k-1}^{k} \frac{\mathrm{d}x}{x+\theta}\,, \tag{4}$$

puisque la fonction $x \mapsto 1/(x+\theta)$ est décroissante. On somme Eq. (4) pour k allant de 0 à $n-1$ et on obtient

$$\int_0^n \frac{\mathrm{d}x}{x+\theta} \leq \sum_{k=0}^{n-1} \frac{1}{k+\theta} \leq \frac{1}{\theta} + \int_0^{n-1} \frac{\mathrm{d}x}{x+\theta}\,.$$

Une primitive de $x \mapsto 1/(x+\theta)$ est $x \mapsto \log(x+\theta)$, on en déduit

$$\log(n+\theta) - \log\theta \leq \sum_{k=0}^{n-1} \frac{1}{k+\theta} \leq \frac{1}{\theta} + \log(n-1+\theta) - \log\theta\,.$$

En soustrayant $\log n$ partout, on arrive à

$$\log\left(1+\frac{\theta}{n}\right) - \log\theta \leq \sum_{k=0}^{n-1} \frac{1}{k+\theta} - \log n \leq \frac{1}{\theta} + \log\left(1+\frac{\theta-1}{n}\right) - \log\theta\,.$$

On en déduit que

$$\left|\sum_{k=0}^{n-1} \frac{1}{k+\theta} - \log n\right| \leq |\log\theta| + \theta + \frac{1}{\theta}\,,$$

et par conséquent

$$\left|\mathbb{E}\left[|B_n|\right] - \theta\log n\right| \leq |\theta\log\theta| + \theta^2 + 1\,.$$

On en déduit que $\hat{\theta}_n$ est asymptotiquement sans biais.

Remarque : il n'y avait pas besoin d'être aussi précis dans les constantes pour avoir tous les points.

II.4. Le même raisonnement qu'à la question II.3. permet d'obtenir

$$\sum_{k=1}^{n-1} \frac{k}{(k+\theta)^2} - \log n = \mathcal{O}\left(1\right)\,.$$

On en déduit que

$$\mathrm{Var}\left(|B_n|\right) - \theta\log n = \mathcal{O}\left(1\right)\,,$$

puis

$$\frac{\mathrm{Var}\left(|B_n|\right)}{\theta\log n} = 1 + o\left(1\right)\,.$$

II.5. On revient à la définition de la convergence en probabilité : soit $\varepsilon > 0$, nous allons montrer que

$$\mathbb{P}\left(\left|\frac{|B_n|}{\log n} - \theta\right| > \varepsilon\right) \longrightarrow 0\,.$$

Après réarrangement, il apparaît qu'on doit contrôler

$$\mathbb{P}\left(\left||B_n| - \theta\log n\right| > \varepsilon\log n\right)\,.$$

Puisqu'on connaît la variance de $|B_n|$, une approche naturelle est faire apparaître le carré de B_n, puis d'utiliser l'inégalité de Markov :

$$\mathbb{P}\left(\left||B_n| - \theta\log n\right| > \varepsilon\log n\right) \leq \frac{\mathbb{E}\left[\left(|B_n| - \theta\log n\right)^2\right]}{\varepsilon^2\log^2 n}\,.$$

On remarque que

$$\mathbb{E}\left[(|B_n| - \theta \log n)^2\right] = \mathrm{Var}\left(|B_n|\right) + \left(\mathbb{E}\left[|B_n|\right] - \theta \log n\right)^2 = \mathcal{O}\left(\log n\right)$$

d'après les questions II.3. et II.4. On en déduit que

$$\mathbb{P}\left(\left||B_n| - \theta \log n\right| > \varepsilon \log n\right) = \mathcal{O}\left(\frac{1}{\log n}\right),$$

et on peut conclure : l'estimateur est consistant.

II.6. Pour pouvoir utiliser Borel-Cantelli et obtenir la convergence forte, il faudrait que la borne obtenue dans la question II.5. soit telle que

$$\sum_{n \geq 1} p_n < +\infty\,.$$

Ce n'est pas le cas ici : $\sum_n \frac{1}{\log n}$ ne converge pas.

III.1. Posons $a = C_n$. Il y a trois possibilités :

1. le citoyen $n+1$ crée son propre parti, et donc $C_{n+1} = a + 1$. Cet événement arrive avec probabilité $\theta/(n+\theta)$;
2. le citoyen $n+1$ rejoint un micro-parti, et donc $C_{n+1} = a - 1$ (le parti rejoint n'est plus un micro-parti). Ceci arrive avec probabilité $a/(n+\theta)$;
3. le citoyen $n+1$ rejoint un parti qui n'est pas un micro-parti, ce qui arrive avec probabilité

$$1 - \frac{a}{n+\theta} - \frac{\theta}{n+\theta} = \frac{n-a}{n+\theta}\,.$$

III.2. D'après la question précédente,

$$\begin{aligned}
\mathbb{E}\left[C_{n+1} \mid C_n = a\right] &= (a+1) \cdot \mathbb{P}\left(C_{n+1} = a+1 \,|\, C_n = a\right) \\
&\quad + a \cdot \mathbb{P}\left(C_{n+1} = a \,|\, C_n = a\right) \\
&\quad + (a-1) \cdot \mathbb{P}\left(C_{n+1} = a-1 \,|\, C_n = a\right) \\
&= (a+1) \cdot \frac{\theta}{n+\theta} + a \cdot \frac{a}{n+\theta} + (a-1) \cdot \frac{n-a}{n+\theta} \\
\mathbb{E}\left[C_{n+1} \mid C_n = a\right] &= \frac{a(n+\theta-1)+\theta}{n+\theta}\,.
\end{aligned}$$

III.3. L'astuce ici est de prendre l'espérance des deux côtés de la dernière ligne de calcul dans la question III.2. En posant $s_n := \mathbb{E}\left[C_n\right]$, on obtient

$$(n+\theta)s_{n+1} = (n+\theta-1)s_n + \theta\,.$$

Posons $u_n := (n-1+\theta)s_n$: c'est une suite récurrente linéaire d'ordre 1 qui vérifie

$$\begin{cases} u_1 & = \theta s_1 = \theta \\ u_{n+1} & = u_n + \theta \quad \forall n \geq 1 \end{cases}$$

On en déduit $u_n = n\theta$, puis immédiatement

$$\mathbb{E}\left[C_n\right] = \frac{n\theta}{n-1+\theta}\,. \tag{5}$$

Remarque : on pouvait avoir tous les points ici sans avoir l'idée de poser u_n et simplement en vérifiant que la formule fournie par l'énoncé satisfaisait la relation de récurrence.

III.4. On remarque que $\mathbb{E}\left[C_n\right] \longrightarrow \theta$. De la même manière que dans la deuxième partie, on pose

$$\hat{\theta}_n := C_n\,.$$

III.5. Pour calculer la variance de C_n, il faut calculer $\mathbb{E}\left[C_n^2\right]$. On est tenté de répéter le raisonnement des questions III.2. et III.3. Calculons tout d'abord l'espérance conditionnelle :

$$\begin{aligned}
\mathbb{E}\left[C_{n+1}^2 \mid C_n = a\right] &= (a+1)^2 \cdot \mathbb{P}\left(C_{n+1} = a+1 \,|\, C_n = a\right) \\
&\quad + a^2 \cdot \mathbb{P}\left(C_{n+1} = a \,|\, C_n = a\right) \\
&\quad + (a-1)^2 \cdot \mathbb{P}\left(C_{n+1} = a-1 \,|\, C_n = a\right) \\
&= (a+1)^2 \cdot \frac{\theta}{n+\theta} + a^2 \cdot \frac{a}{n+\theta} + (a-1)^2 \cdot \frac{n-a}{n+\theta} \\
\mathbb{E}\left[C_{n+1}^2 \mid C_n = a\right] &= \frac{a^2(\theta+n-2) + a(2\theta+1) + \theta}{n+\theta}\,.
\end{aligned}$$

Posons $t_n := \mathbb{E}\left[C_n^2\right]$. En prenant l'espérance des deux côtés dans la dernière ligne, on obtient

$$t_{n+1} = \frac{t_n(\theta+n-2) + n\theta(2\theta+1)/(n-1+\theta) + \theta}{n+\theta}\,,$$

où l'on a utilisé le résultat de la question III.2. Après quelques simplifications, on obtient

$$(n+\theta)(n+\theta-1)t_{n+1} = (n+\theta-1)(n+\theta-2)t_n + n\theta(2\theta-1) + \theta(n+\theta-1)\,. \tag{6}$$

Par analogie avec la question III.3., on pose

$$v_n := (n+\theta-1)(n+\theta-2)t_n$$

Eq. (6) devient alors

$$v_{n+1} = v_n + n(2\theta^2+2\theta) + \theta^2 - \theta\,.$$

Ainsi v est une suite récurrente linéaire d'ordre 2, avec des termes non constants mais de petit degré. Comme $v_1 = \theta^2 - \theta$, on en déduit

$$\begin{aligned}
v_n &= \frac{n(n-1)}{2} \cdot (2\theta^2+2\theta) + n \cdot (\theta^2-\theta) \\
&= n\theta[(n-1)(\theta+1) + \theta - 1] \\
v_n &= n\theta(n\theta+n-2)
\end{aligned}$$

où l'on a utilisé que $\sum_{k=1}^{n} = n(n+1)/2$ pour passer de la première ligne à la deuxième. On en déduit que

$$\mathbb{E}\left[C_n^2\right] = \frac{n\theta(n\theta+n-2)}{(n+\theta-2)(n+\theta-1)}\,. \tag{7}$$

Il suffit maintenant de soustraire (le carré de) Eq. (5) à Eq. (7) pour obtenir la variance de C_n. Après un long calcul, on obtient

$$\mathrm{Var}\left(C_n\right) = \frac{n(n-1)(n-2+2\theta)\theta}{(n+\theta-2)(n+\theta-1)^2}\,.$$

III.6. Les deux estimateurs sont tous les deux asymptotiquement sans biais, et nous n'avons pas d'information sur la convergence du second. On ne peut donc comparer que la variance des deux estimateurs. D'après la question II.4., $\mathrm{Var}\left(|B_n|\right) \sim \theta \log n$, on en déduit $\mathrm{Var}(\hat{\theta}_n) \sim \frac{\theta}{\log n}$. C'est à comparer avec $\mathrm{Var}\left(C_n\right)$, qui d'après la question III.5. vérifie $\mathrm{Var}\left(C_n\right) \sim \theta$. On préfère utiliser l'estimateur avec la plus petite variance, c'est-à-dire $\hat{\theta}_n$.