Comportement asymptotique du nombre de valeurs distinctes dans un échantillon de cassage de bâton géométrique.

Doron Israel
doron.israel@dauphine.eu

Emmanuel Memmi
emmanuel.memmi@ens.fr

Résumé

Cet article de De Blasi, Mena et Prünster [1] s'intéresse au processus de cassage de bâton géométrique, et à l'estimation asymptotique de K_n, le nombre de valeurs distinctes pour des échantillons définis suivant ce processus.
Le résultat principal est le théorème 1, qui donne un développement asymptotique de $E(K_n)$ en fonction de la répartition des basses fréquences du processus. L'article considère ensuite plusieurs applications de ce théorème. Après avoir présenté les grands résultats de l'article, nous réalisons quelques simulations en Python pour illustrer les conséquences de ce théorème. [1]

1 Définir le cassage de bâton géométrique

Le processus de cassage de bâton géométrique ("geometric stick breaking"), peut être défini de la manière suivante

1. On prend un prior sur p sur $]0, 1[$
2. On définit les fréquences géométriques $w_j = p(1-p)^{j-1}$ pour $j \in \mathbb{N}_{\geq 1}$ (il est à noter qu'elles sont décroissantes)
3. On peut alors définir une distribution de probabilités discrète et *aléatoire* :

$$\tilde{p}(dx) = \sum_{j \geq 1} w_j \delta_{x_j}(dx).$$

où les x_j sont des éléments d'un ensemble polonais $\mathbb{X}$.

2 Modélisation du problème

On cherche à étudier le nombre de valeurs distinctes atteintes pour un n-échantillon du processus défini précédemment. Pour ce faire, il est important de choisir la bonne modélisation.

Considérons d'abord le cas p fixé.

Une première modélisation est que ce processus peut être vu comme l'expérience de lancers de boules de manière indépendantes dans une série infinie fixée de boîtes, avec une probabilité w_j de tomber dans la $j^{\text{ème}}$ boîte. Lorsque n boules sont lancées, leur emplacement est capturé par le vecteur $X_n = (X_{n,j})_{j \geq 1}$ où $X_{n,j}$ représente le nombre de boules dans la boîte j après n lancers. Notre objet d'intérêt est donc $K_n := \sum_{j \geq 1} \mathbf{1}_{X_{n,j} > 0}$ On déduit facilement que :

$$\mathbb{E}(K_n) = \sum_{j \geq 1} 1 - \mathbb{P}(X_{n,j} = 0) = \sum_{j \geq 1} (1 - (1 - w_j)^n)$$

1. code disponible sur le Github du projet.

Soutenance de Bayesian Non Parametrics, Dauphine, mars 2021.

Il est malheureusement difficile de travailler avec cette définition de K_n car les $(\mathbf{1}_{X_{n,j}>0})_{j\geq 1}$ ne sont pas indépendantes.

Une deuxième modélisation consiste à considérer que l'on lance les boules à chaque saut d'un processus de Poisson $P(t)_{t\geq 0}$ de paramètre 1 qui est indépendant de $(X_n)_{n\geq 1}$. Si on considère maintenant le processus $X_j(t)_{t\geq 0}$ défini comme le nombre de boules dans la boîte j au temps t alors $X_j(t)$ est un processus de Poisson de paramètre w_j.

Montrons qu'alors $\forall t \geq 0 \;\; X_j(t)$ suit une loi de Poisson de paramètre $t\, w_j$.

Soit $t \geq 0$. Pour tout $k \in \mathbb{N}$ on a :

$$\begin{aligned}
\mathbb{P}(X_j(j) = k) &= \sum_{n\geq 0} \mathbb{P}(X_j(t) = k | P(t) = n) \times \mathbb{P}(P(t) = n) \\
&= \sum_{n\geq k} \binom{n}{k} w_j^k (1-w_j)^{n-k} \frac{t^n}{n!} e^{-t} = \frac{(t\,wj)^k}{k!} e^{-t} \sum_{n\geq k} \frac{((1-w_j)t)^{n-k}}{(n-k)!} \\
&= \frac{(t\,wj)^k}{k!} e^{-t} e^{t(1-wj)} = \frac{(t\,wj)^k}{k!} e^{-twj}
\end{aligned}$$

On définit alors $K(t) := K_{P(t)}$, c'est à dire comme le nombre de boîtes distinctes dans lesquelles ont été lancées les $P(t)$ boules au temps t. On a alors :

$$\Phi(t) := \mathbb{E}(K(t)) = \sum_{j\geq 1} (1 - \mathbb{P}(X_j(t) = 0)) = \sum_{j\geq 1} (1 - e^{-tw_j})$$

En considérant ensuite la mesure de comptage sur les fréquences $\nu(dx) = \sum_{j\geq 1} \delta_{w_j}(dx)$ on peut ré-écrire $\Phi(t)$ et $\mathbb{E}(K_n)$ comme ceci :

$$\begin{aligned}
\mathbb{E}(K_n) &= \int_0^1 (1 - (1-x)^n)\, \nu(\mathrm{d}x) \\
\Phi(t) &= \int_0^1 (1 - e^{tx})\, \nu(\mathrm{d}x) = t \int_0^1 e^{-tx} \overrightarrow{\nu}(\mathrm{d}x)
\end{aligned}$$

où $\overrightarrow{\nu}(x) = \nu([x,1])$ représente le nombre de fréquences plus grandes que x — la deuxième égalité de la deuxième ligne provient d'une simple intégration par parties.

Un dernier élément afin de justifier cette seconde modélisation pour l'étude de K_n, pour lequel on peut trouver une preuve dans [3], lemme 1 :

$$|\mathbb{E}(K_n) - \Phi(n)| \leq \frac{2}{n} \Phi(n) \to 0$$

Dans le cas ou p est aléatoire, il suffit de remplacer la mesure de comptage $\nu(x)$ par sa moyenne selon la loi de p, on obtient ainsi que :

$$\overrightarrow{\nu}(x) = \int_0^1 \overrightarrow{\nu}(x,p)\pi(p)dp$$

Cette partie sur la modélisation du problème a été introduite afin de mieux comprendre les éléments utilisés dans la suite pour étudier le comportement asymptotique de K_n.

3 Théorème fondamental

Théorème 1. *S'il existe $c \geq 0$ et une fonction ℓ tels que $\frac{\overrightarrow{\nu}(\lambda x) - \overrightarrow{\nu}(x)}{l(x)} \xrightarrow[x\to 0]{} c \log \lambda$ pour tout $\lambda > 0$,*

avec $\ell(\lambda x)/\ell(x) \xrightarrow[x\to 0]{} 1$ pour tout $\lambda > 0$ (on dit que ℓ est une fonction variant lentement en 0),

alors

$$E(K_n) = \overrightarrow{\nu}(1/n) - c\gamma\ell(1/n) + o(l(1/n)) \;\; \textit{lorsque } n \to \infty$$

avec γ la constante d'Euler.

Démonstration. Notons d'abord que $E(K_n) \sim \overrightarrow{\nu}(1/n)$ lorsque $n \to \infty$: cela vient du fait que $E(K_n) \sim \Phi(n)$ ([3]), et que $\Phi(n) \sim \overrightarrow{\nu}(1/n)$ lorsque $n \to \infty$ (conséquence du théorème taubérien [4].)

On écrit ensuite $E(K_n) = \overrightarrow{\nu}(1/n) + \frac{\Phi(n)-\overrightarrow{\nu}(1/n)}{\ell(1/n)}\ell(1/n) + E(K_n) - \Phi(n)$.

Comme $E(K_n) - \Phi(n) = o(l(1/n))$, il suffit d'estimer la quantité $\big(\Phi(n) - \overrightarrow{\nu}(1/n)\big)/\ell(1/n)$:

$$\begin{aligned}\frac{\Phi(1/x) - \overrightarrow{\nu}(x)}{\ell(x)} &= \frac{1}{\ell(x)}\left[\int_0^\infty \frac{1}{x}e^{-y/x}\overrightarrow{\nu}(y)dy - \int_0^\infty \overrightarrow{\nu}(x)e^{-\lambda}d\lambda\right] \\ &= \int_0^\infty \frac{\overrightarrow{\nu}(\lambda x) - \overrightarrow{\nu}(x)}{\ell(x)}e^{-\lambda}d\lambda\end{aligned}$$

Or, par convergence dominée, cette dernière quantité converge quand $x \to 0$ vers

$$\int_0^\infty c(\log\lambda)e^{-\lambda}d\lambda = c\Gamma'(1) = -c\gamma$$

□

4 Applications et simulations

4.1 Un premier exemple

Exemple 1. On fixe p dans $]0,1[$ — autrement dit, on prend un prior dirac en p.

Notons qu'on peut expliciter $\overrightarrow{\nu}(x,p)$: il s'agit en effet du plus grand entier j tel que $p(1-p)^{j-1} \geq x$ et on peut résoudre en j l'équation $p(1-p)^{j-1} = x$.

Alors $\overrightarrow{\nu}(x,p) = \left\lfloor \frac{\log(x/p)}{|\log(1-p)|} + 1 \right\rfloor 1_{x\leq p} \underset{x\to 0}{\sim} \frac{\log(x)}{\log(1-p)}$ et $\overrightarrow{\nu}$ vérifie les hypothèses du théorème avec ℓ constante égale à 1 et $c = \frac{1}{\log(1-p)}$.

Le théorème donne alors

$$E(K_n) = \left\lfloor \frac{\log(np)}{|\log(1-p)|} \right\rfloor + 1 + \frac{\gamma}{|\log(1-p)|} + o(1) \quad \text{lorsque } n \to \infty.$$

Illustration 1.

Nous avons illustré ce résultat en ne gardant que le premier terme du développement : notons qu'à partir de la formule ci-dessus on a en particulier

$$E(K_n) \sim \frac{\log(n)}{|\log(1-p)|}$$

lorsque $n \to \infty$.
On constate que le simple équivalent fonctionne extrêmement bien en pratique !

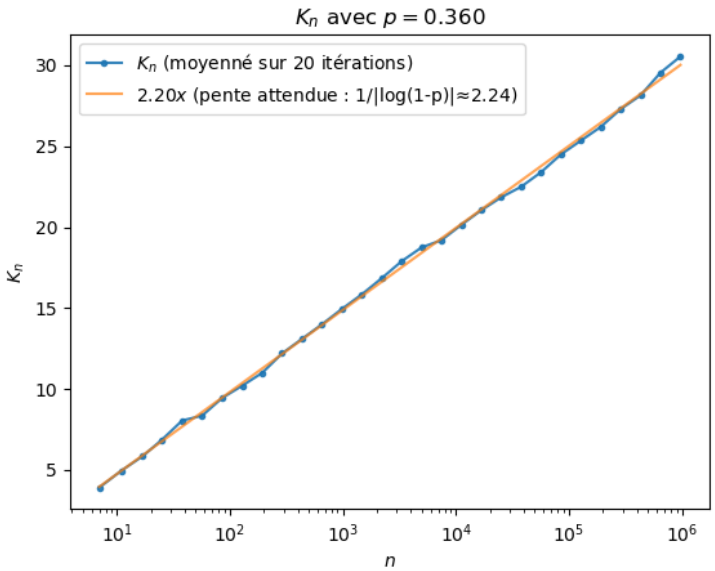


FIGURE 1 – Illustration de l'exemple 1.

4.2 Proposition 1

Proposition 1. *On prend p suivant une loi uniforme sur $]0,1[$. Alors on a*

$$\overrightarrow{\nu}(x) = \frac{1}{2}(\log x)^2 + \gamma \log x + O(1), \quad x \to 0$$

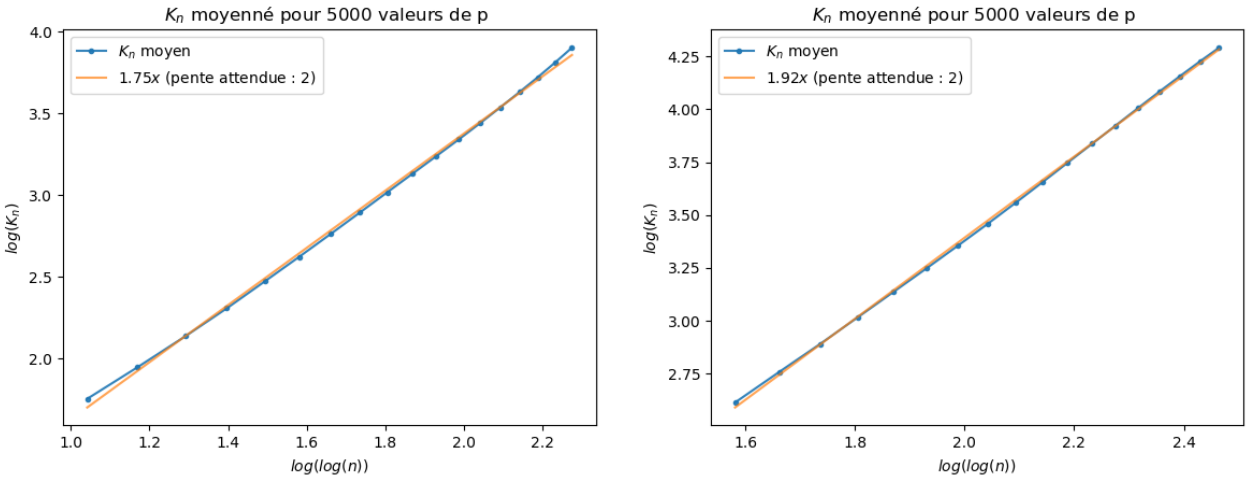
$$E(K_n) = \frac{1}{2}(\log n)^2 + o(\log n), \quad n \to +\infty$$

Illustration 2. Notons que la proposition implique que

$$\log E(K_n) \sim 2 \log\log n$$

On trace donc $\log E(K_n)$ en fonction de $\log\log n$ et on s'attend à y voir une droite linéaire de pente 2.

Ici, l'estimation fonctionne un peu moins bien pour de « faibles » valeurs de n : on trouve systématiquement une pente aux alentours de 1.8 quand on se cantonne à des n en dessous de 10^5. L'estimation fonctionne mieux pour des valeurs élevées de n (mais les simulations prennent assez longtemps).



(a) Illustration de l'exemple 1 pour n entre 15 et 61000.

(b) Illustration de l'exemple 1 pour des n entre 400 et 200000.

FIGURE 2 – Simulations pour l'exemple 1.

4.3 Théorème 2

Théorème 2. *Soit p suivant la même loi que e^{-X} où $X \sim$ Gamma$(m+1, m)$ avec $m \in \mathbb{N}^*$. On a :*

$$\overrightarrow{\nu}(x) = \frac{\log(\frac{1}{x})^{m+2}}{m+2!} + O((\log(x))^m), \quad x \to 0$$

$$\mathbb{E}(K_n) = \frac{(\log n)^{m+2}}{m+2!} + \gamma \frac{(\log n)^{m+1}}{m+1!} + o((\log n)^{m+1}), \quad n \to +\infty$$

Illustration 3.

On s'attend cette fois à observer

$$\log E(K_n) \sim (m+2) \log\log n$$

On peut tester pour n'importe quelle valeur de m. On constate que pour de plus grandes valeurs de m, on a un peu plus de mal à observer le taux théorique pour les valeurs de n considérées.

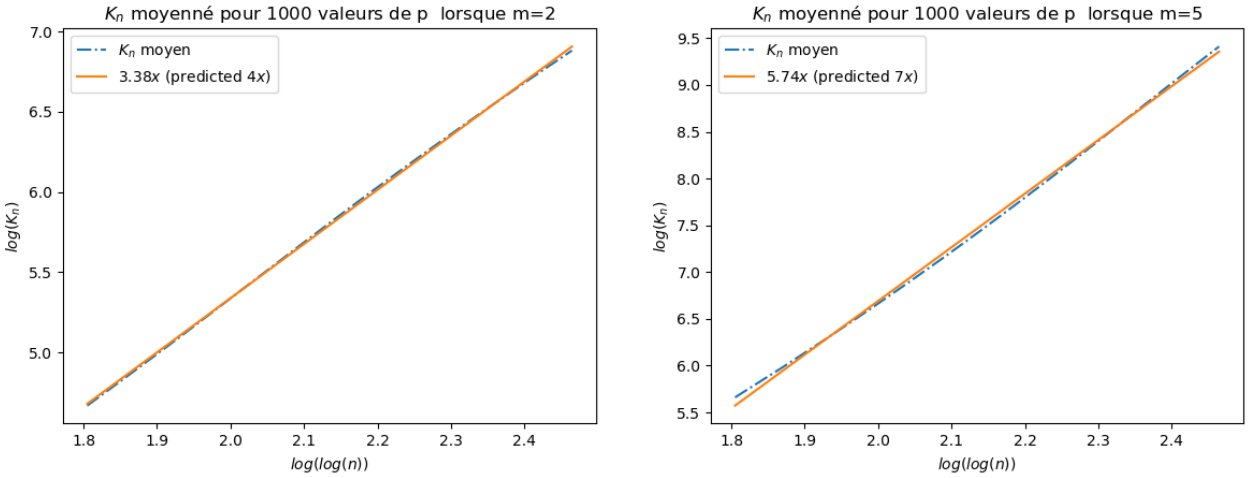


FIGURE 3 – Simulations pour l'exemple 1.

4.4 Proposition 2

On peut ensuite étendre le résultat à des séquences de poids construits d'une différente manière. Considérons :

$$w_j(p) = \sum_{r \geq j} \frac{\phi(r;p)}{r}, \quad \forall j \in \mathbb{N}^*$$

Où $\phi(r;p)$ est une fonction de probabilité sur $r \in \mathbb{N}^*$ dépendant d'un paramètre p. Ainsi définie, $(w_j)_{j\geq 1}$ est bien une suite décroissante de somme 1. Un type de fonction ϕ intéressante est défini par :

$$\phi(r;p) = \phi(r;s,p) = \binom{r+s-2}{r-1} p^s (1-p)^{r-1}, \quad r \in \mathbb{N}^*$$

Ceci correspond à la binomiale négative décalée de 1. Pour $s = 2$ on retrouve la séquence de poids géométriques $w_j = p(1-p)^{j-1}$. Pour la proposition suivante on s'intéressera au cas $s = 3$ qui correspond aux poids $w_j = p(1-p)^{j-1}\frac{1+jp}{2}$ pour $j \in \mathbb{N}^*$.

Proposition 2. *Soit p suivant la loi uniforme sur $[0,1]$. Alors :*

$$\overrightarrow{\nu}(x) = \frac{1}{2}\left(\log\frac{1}{x}\right)^2 + \log\log\frac{1}{x} - (1+\log 2)\log\frac{1}{x} + O\left(\log\log\frac{1}{x}\right), \quad x \to 0$$
$$\mathbb{E}(K_n) = \frac{1}{2}(\log n)^2 + \log\log n - (1+\log 2)\log n + o(\log n), \quad n \to +\infty$$

Illustration 4. Encore une fois, la proposition implique que

$$\log E(K_n) \sim 2\log\log n$$

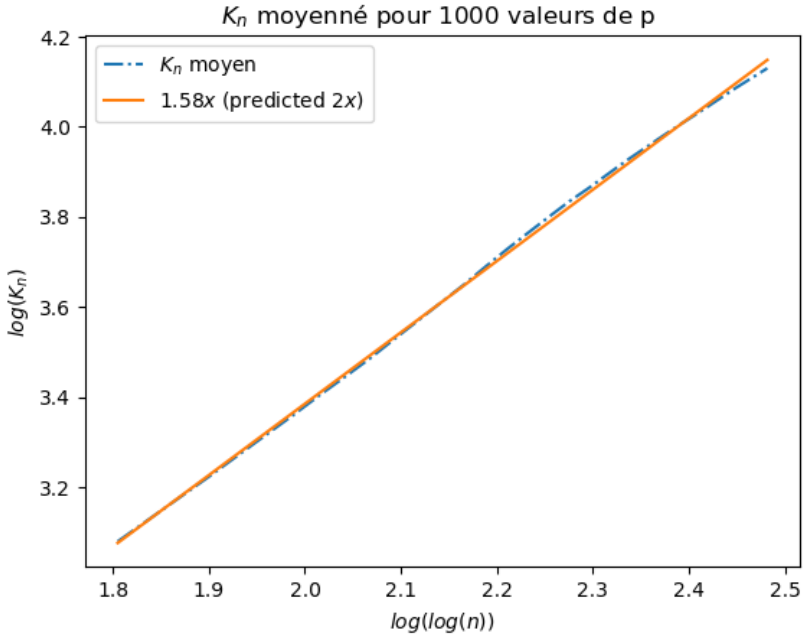


FIGURE 4 – Illustration de la proposition 2.

5 Quelques coquilles dans l'article

Nous avons relevé quelques coquilles dans l'article. Bien que celles-ci n'aient pas d'influence sur les résultats, nous les indiquons ici :

- à plusieurs reprises, il est écrit F à la place de f. Page 4 lors de la Notation : « Let $F(x)$ [...] the fractional integral of order α of $F(x)$ » : tous les F devraient être des f. De même, en page 7 au début de la preuve de la proposition 1 : « is the fractional integral of order 1 of F ».
- dans la preuve de la proposition 1 (page 7), à la première ligne du développement asymptotique de $m(\lambda x) - m(x)$, on devrait avoir un signe $-$ et non un $+$ devant le terme $\gamma(\log(\lambda x) - \log(x))$. Par chance ce terme est un $O(1)$ donc le signe n'a pas d'importance pour le résultat final.

Bibliographie

[1] Pierpaolo De Blasi and Ramsés H. Mena and Igor Prünster, (2021). Asymptotic behavior of the number of distinct values in a sample from the geometric stick-breaking process.

[2] Fuentes-García, R., Mena, R. H., and Walker, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics - Simulation and Computation*, 39(4) :669–682.

[3] Gnedin, A., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes : general asymptotics and power laws. *Probability Surveys*, 4 :146–171.

[4] Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*. Cambridge University Press.