The coupon collector's problem: from moments to asymptotic analysis

We address here a classical probability problem called the coupon collector's problem. This problem, studied for instance by Laplace in the beginning of the 19th century, can be phrased in many equivalent ways. In modern and commercial terms, the problem is that of a person who is collecting coupons (or toys) offered with products, say boxes of a brand of cereals. If each box contains a coupon, if the coupons are uniformly distributed in the boxes, and if there are N different types of coupons (or toys), then the problem consists in finding the expected number of boxes that need to be bought to collect all N coupons, or more generally the distribution of that number of boxes. This problem has been addressed in many ways and the goal of the below problem set is to walk through the different approaches.

Preliminaries: Harmonic series

Let us introduce the Harmonic series:

$$\forall n \in \mathbb{N}^*, H_n = \sum_{k=1}^{n} \frac{1}{k}.$$

1. Show that the series $\sum_{k\geq 2} \left(\frac{1}{k} - \log\left(\frac{k}{k-1}\right)\right)$ converges.
2. Show that

$$H_n = \log(n) + \gamma + o_{n\to+\infty}(1), \qquad \text{where } \gamma = 1 + \sum_{k=2}^{\infty} \left(\frac{1}{k} - \log\left(\frac{k}{k-1}\right)\right).$$

Remark: the constant $\gamma \simeq 0.57721$ is called the Euler-Mascheroni's constant.

A first approach for the mean and the variance

Let us denote by N the number of different coupons.

For $n \in \{1, \dots, N\}$, let us denote by T_n the random variable corresponding to the number of boxes bought when the collector obtains its nth different coupon.

1. What is T_1?
2. Explain why the random variables $T_2 - T_1, T_3 - T_2, \dots, T_N - T_{N-1}$ are independent (no math is required). For $n \in \{2, \dots, N\}$, what is the distribution of $T_n - T_{n-1}$?
3. Deduce that

$$\forall n \in \{2, \dots, N\}, \mathbb{E}[T_n - T_{n-1}] = \frac{N}{N-(n-1)}.$$

4. Conclude that the expected number of boxes that need to be bought to collect all N coupons is $\mathbb{E}[T_N] = NH_N$. Give an asymptotic expansion of $\mathbb{E}[T_N]$ when $N \to +\infty$.
5. Compute $\mathbb{E}[T_{100}]$ and compare the result with the approximation obtained with the above asymptotic expansion.
6. Show that

$$\mathbb{V}[T_N] = \sum_{n=1}^{N} \frac{\frac{n-1}{N}}{\left(1 - \frac{n-1}{N}\right)^2}.$$

7. Give an asymptotic expansion of $\mathbb{V}[T_N]$ when $N \to +\infty$.
8. Compute $\mathbb{V}[T_{100}]$ and compare the result with the approximation obtained with the above asymptotic expansion.

Towards the distribution of T_N

We now consider that the coupons are numbered from 1 to N. For $n \in \{1, \dots, N\}$, we denote by X_n the number of boxes bought when the coupon n is obtained for the first time.

9. Show that $\forall k \in \mathbb{N}^*, \mathbb{P}(T_N \geq k) = \mathbb{P}\left(\bigcup_{n=1}^{N} \{X_n \geq k\}\right)$.
10. Use Poincaré's formula to deduce that

$$\forall k \in \mathbb{N}^*, \mathbb{P}(T_N \geq k) = \sum_{j=1}^{N} (-1)^{j-1} \sum_{1 \leq i_1 < \ldots < i_j \leq N} \mathbb{P}(\min_{1 \leq l \leq j} X_{i_l} \geq k).$$

11. Show that, for all $j \in \{1, \dots, N\}$, and all $1 \leq i_1 < \ldots < i_j \leq N$, the random variable $\min_{1 \leq l \leq j} X_{i_l}$ has a geometric distribution with parameter $\frac{j}{N}$.
12. Deduce that $\forall k \in \mathbb{N}^*, \mathbb{P}(T_N \geq k) = \sum_{j=1}^{N} (-1)^{j-1} C_N^j \left(1 - \frac{j}{N}\right)^{k-1}$.
13. Conclude that $\forall k \in \mathbb{N}^*, \mathbb{P}(T_N = k) = \sum_{j=1}^{N} (-1)^{j-1} C_{N-1}^{j-1} \left(1 - \frac{j}{N}\right)^{k-1}$
14. Use that expression to show that $\mathbb{E}[T_N] = N \sum_{j=1}^{N} \frac{(-1)^{j-1}}{j} C_N^j$.
15. (Bonus) Show that the two expressions for the expected value of T_N are equal.

Asymptotic analysis: preliminaries on representations of the Γ function

We define for $z \in \mathcal{H} = \{z \in \mathbb{C} | \Re(z) > 0\}$ (complex numbers with positive real part) the Γ function by:

$$z \mapsto \int_0^{+\infty} t^{z-1} e^{-t} dt.$$

16. Justify that Γ is well defined.
17. Show that $\forall z \in \mathcal{H}, \Gamma(z) = \lim_{n \to +\infty} \int_0^n t^{z-1} \left(1 - \frac{t}{n}\right)^n dt$
18. Using n integrations by parts, show that

$$\forall n \in \mathbb{N}^*, \int_0^n t^{z-1} \left(1 - \frac{t}{n}\right)^n dt = \frac{n! n^z}{\prod_{k=0}^{n} (z+k)}.$$

19. Deduce that

$$\forall z \in \mathcal{H}, \Gamma(z) = \frac{1}{z} \prod_{k=1}^{\infty} \frac{\left(1 + \frac{1}{k}\right)^z}{1 + \frac{z}{k}}.$$

20. Using the results of the first part of the problem, show that

$$\forall z \in \mathcal{H}, \Gamma(z) = \frac{e^{-\gamma z}}{z} \prod_{k=1}^{\infty} \frac{e^{\frac{z}{k}}}{1 + \frac{z}{k}}.$$

Asymptotic analysis: convergence towards a Gumbel distribution

21. Show that the characteristic function of a geometric random variable with parameter $p \in (0, 1)$ is

$$\phi_p : \xi \in \mathbb{R} \mapsto \frac{1}{1 + \frac{1}{p}(e^{-i\xi} - 1)}.$$

22. Show that the characteristic function of T_N is

$$\xi \mapsto \prod_{k=1}^{N} \phi_{1 - \frac{k-1}{N}}(\xi) = \prod_{k=1}^{N} \phi_{\frac{k}{N}}(\xi).$$

23. Deduce that the characteristic function of $\frac{T_N - \mathbb{E}[T_N]}{N}$ is

$$\psi_N : \xi \mapsto e^{-i\xi H_N} \prod_{k=1}^{N} \phi_{\frac{k}{N}}\left(\frac{\xi}{N}\right) = \prod_{k=1}^{N} \frac{e^{-i\frac{\xi}{k}}}{\left(1 + \frac{N}{k}(e^{-i\frac{\xi}{N}} - 1)\right)}.$$

24. Show that

$$\Gamma(1 - i\xi) = e^{i\gamma\xi} \prod_{k=1}^{\infty} \frac{e^{-i\frac{\xi}{k}}}{1 - i\frac{\xi}{k}}.$$

25. (Technical) Prove that $\lim_{N\to+\infty} \prod_{k=1}^{N} \frac{1-i\frac{\xi}{k}}{\left(1+\frac{N}{k}(e^{-i\frac{\xi}{N}}-1)\right)} = 1$.

26. Conclude that $\lim_{N\to+\infty} \psi_N(\xi) = e^{-i\gamma\xi}\Gamma(1 - i\xi)$.

27. Let Z be a random variable with cumulative distribution function $F : x \mapsto e^{-e^{-(x+\gamma)}}$ (Gumbel distribution). Compute the characteristic function of Z. Deduce that Z_N converges in distribution towards Z when $N \to +\infty$.

28. Conclude that

$$\forall x \in \mathbb{R}, \lim_{N\to+\infty} \mathbb{P}(T_N > N\log(N) + Nx) = 1 - e^{-e^{-x}}.$$