1 Problem formulation

The question is simple: suppose that \( m \) balls are thrown independently and uniformly at random into \( n \) bins. How large need \( m \) be in order to ensure that the \( n \) bins are all occupied by at least one ball? This problem was first studied by Laplace in 1812. Since then, many variations of the problem have been studied, with interesting applications in various areas (including the one of estimating how much money you need to spend, once every four years, in order to complete an empty book with 682 stickers\(^1\)).

2 Expected behaviour

For \( k \in \{1, \ldots, n\} \), define \( T_k \) to be the first time (i.e., the smallest value of \( m \)) such that \( k \) bins are occupied. We can compute the expected value of \( T_n \) by the following reasoning. Let \( X_1 = T_1 = 1 \) and \( X_k = T_k - T_{k-1} \) for \( k \in \{2, \ldots, n\} \) (in words, \( X_k \) is the waiting time for a new bin to be reached after \( k-1 \) of them have already been reached). Then

\[
P(\{X_k = \ell\}) = \left( \frac{k-1}{n} \right)^{\ell-1} \left( 1 - \frac{k-1}{n} \right) \quad \text{for } \ell \geq 1
\]

i.e., \( X_k \) is a geometric random variable of parameter \( p_k = 1 - \frac{k-1}{n} = \frac{n-k+1}{n} \). Such a random variable has expectation

\[
E(X_k) = \frac{1}{p_k} = \frac{n}{n-k+1}
\]

logically translating the fact that the average time duration for a new bin to be reached increases with \( k \). We finally obtain

\[
E(T_n) = \sum_{k=1}^{n} E(X_k) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k} \simeq n \log n
\]

as \( n \) gets large (we spare you here the additional Euler constant that refines this approximation).

3 Threshold phenomenon

The result below (due to Erdős and Rényi) shows that \( m = n \log n \) is not only an average behaviour, but also the precise threshold before which the coupon collection is not complete with high probability and after which it is complete with high probability.

**Proposition.** For \( t \in \mathbb{R} \), we have

\[
\lim_{n \to \infty} P(\{T_n \leq n \log n + nt\}) = \exp(-e^{-t})
\]

In other words, this is saying that the sequence of random variables \( (G_n, n \geq 1) \) defined as

\[
G_n = \frac{T_n - n \log n}{n}
\]

\(^1\)For those interested, the answer was around 1’000 Swiss Francs for the last edition!
converges in distribution as $n \to \infty$ towards the random variable $G$ with cdf

$$F_G(t) = \exp(-e^{-t}), \quad t \in \mathbb{R}$$

also known as the standard Gumbel distribution.

**Remark.** The implications of the above result are the following:
- If $t$ is large and positive, then $P(\{T_n \leq n \log n + nt\}) \simeq \exp(-e^{-t}) \simeq \exp(0) \simeq 1$.
- If $t$ is large and negative, then $P(\{T_n \leq n \log n + nt\}) \simeq \exp(-e^{-t}) \simeq \exp(-\infty) \simeq 0$.

which proves the claim made at the beginning of this section.

**Proof sketch.** We do not provide below a complete proof of the above proposition, but just an approximation argument. For $m \geq 1$ and $i \in \{1, \ldots, n\}$, define

$$E_{im} = \{\text{bin } i \text{ is still empty after } m \text{ throws}\}$$

Considering $m = \lceil n \log n + tn \rceil$ with $t \in \mathbb{R}$ fixed, we obtain as $n$ gets large:

$$P(E_{im}) = \left(1 - \frac{1}{n}\right)^m \simeq \exp\left(-\frac{m}{n}\right) \simeq \exp(-\log n - t) = \frac{e^{-t}}{n}, \quad \forall i \in \{1, \ldots, n\} \quad (1)$$

Besides, the events $E_{1m}, \ldots, E_{nm}$ are “approximately independent” in the following sense$^2$: for every $i \in \{1, \ldots, m\}$ and $J \subset \{1, \ldots, n\}$ such that $J \cap \{i\} = \emptyset$ and $|J| = k$ with $k$ fixed, we have

$$P(E_{im} | \cap_{j \in J} E_{jm}) \simeq P(E_{im}) \quad \text{as } n \to \infty \quad \text{(and } m = \lceil n \log n + tn \rceil) \quad (2)$$

Indeed:

$$P(E_{im} | \cap_{j \in J} E_{jm}) = \frac{P(E_{im} \cap (\cap_{j \in J} E_{jm}))}{P(\cap_{j \in J} E_{jm})} = \frac{(1 - (k + 1)/n)^m}{(1 - k/n)^m} = \left(\frac{n - k - 1}{n - k}\right)^m = \left(1 - \frac{1}{n - k}\right)^m$$

$$\simeq \exp\left(-\frac{m}{n - k}\right) \simeq \exp\left(-\frac{m}{n}\right) \simeq P(E_{im})$$

for $k$ fixed and $n$ large (with $m = \lceil n \log n + tn \rceil$). Using the above approximations (1) and (2), we obtain

$$P(\{T_n \leq m\}) = P(\{\text{all bins are occupied after } m \text{ throws}\})$$

$$= P\left(\bigcap_{i=1}^{n} E_{im}^c\right) \simeq \prod_{i=1}^{n} P(E_{im}^c) \simeq \left(1 - \frac{e^{-t}}{n}\right)^n \simeq \exp(-e^{-t})$$

which “proves” the claim (for a full proof, the inclusion-exclusion principle is needed, as well as the Bonferroni inequalities). □

$^2$Note that they cannot be all independent, because if for example the first $n - 1$ bins are empty after $m$ throws, then the last one can’t be.