The coupon collector problem

Suppose you throw m balls independently and uniformly at random into n bins; how large need m be to ensure that each bin contains at least one ball?

Expected behaviour

Define $T_k = \text{the first time } k \text{ bins are occupied } \implies k \leq n$

How does $T_n$ behave with n?
Define

\[ X_1 = T_1 = 1, \quad X_k = T_k - T_{k-1} \quad 2 \leq k \leq n \quad T_n = \sum_{k=1}^{n} X_k \]

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

so \( X_k \sim \text{Geom}(p_k) \), \( p_k = 1 - \frac{k-1}{n} = \frac{n-k+1}{n} \)

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

\[ E(T_n) = \sum_{k=1}^{n} E(X_k) = n \cdot \sum_{k=1}^{n} \frac{1}{n-k+1} = n \cdot \sum_{k'=1}^{n} \frac{1}{k'} \approx n \cdot \log n \]
Proposition (Erdős–Renyi)

\[ \forall t \in \mathbb{R}, \lim_{n \to \infty} P(\xi_{T_n} \leq u \log n + nt) = \exp(-e^{-t}) \]

Equivalently: \[ G_n = \frac{T_n - u \log n}{u} \quad n \geq 1 \]

\[ G_n \xrightarrow{n \to \infty} G \quad \text{where} \quad F_G(t) = \exp(-e^{-t}) \quad t \in \mathbb{R} \]

- if \( t \) is large & positive, then \( P(\xi_{T_n} \leq u \log n + nt) \approx \exp(-e^{-t}) \approx 1 \)

- if \( t \) is large & negative, then \( u \approx \exp(-e^{-t}) \approx 0 \)
Illustration:

\[ P(\exists T_i \leq k) \uparrow 1 \]

Proof:

\[ E_i^m = \{ \text{bin } i \text{ is still empty after } m \text{ throws} \} \quad 1 \leq i \leq n \]

\[ m = n \log n + n t \]

\[ P(\{ T_i \leq m \}) = P(\bigcap_{i=1}^{n} E_i^m) \]

\[ P(E_i^m) = (1 - \frac{1}{n})^m \approx \exp(-\frac{m}{n}) = \exp(-\log n - t) = \frac{e^{-t}}{n} \]

\[ \forall 1 \leq i \leq n \]
Besides, the events $E_{1m} \ldots E_{nm}$ are "approximately independent":

$$P(E_{jm} \mid \cap_{j \in J} E_{jn}) \approx P(E_{jm})$$

where $\cap_{j \in J} E_{jn} = \emptyset$, $|J| = k$, $i \in J$,

$$P(E_{jm} \cap (\cap_{j \notin J} E_{jn})) = \frac{P(E_{jm})}{P(\cap_{j \notin J} E_{jn})} = \left(1 - \frac{k+1}{n}\right)^m = \left(\frac{n-k+1}{n-k}\right)^m = \left(1 - \frac{1}{n-k}\right)^m \quad \text{fixed}$$

$$\cap (1 - \frac{1}{n})^m \approx \exp(-\frac{m}{n}) = \frac{e^{-t}}{n} \approx P(E_{jm})$$

$$P(\{T_n \leq m\}) = P(\cap_{i=1}^n E_{im}^c) \approx \prod_{i=1}^n P(E_{im}^c)$$

$$m = n \log n + n t$$

$$\approx \prod_{i=1}^n \left(1 - \frac{e^{-t}}{n}\right) = \left(1 - \frac{e^{-t}}{n}\right)^n \xrightarrow{n \to \infty} \exp(-e^{-t}) \quad \forall t \in \mathbb{R}$$
Moments and Carleman’s Theorem

Def: Let $X$ be a r.v. & $k \geq 0$. If $E(|X|^k) < \infty$, we say that the moment of order $k$ of $X$ is finite & we define $m_k = E(X^k)$.

Ex: $X \sim N(0, 1)$: $E(|X|^k) < \infty \quad \forall k \geq 0$

$X \sim \text{Cauchy}(1)$: $E(|X|^k) = \infty \quad \forall k \geq 1$

Note: $m_0 = E(X^0) = 1 \quad \forall X$

Proposition: if $E(|X|^k) < \infty$, then $E(|X|^l) < \infty \quad \forall 0 \leq l \leq k$

Proof: $E(|X|^l) = E((|X|^k)^{l/k}) \leq \left( E(|X|^k) \right)^{l/k}$

$f(x) = x^{l/k}$ concave

Jensen
Question 1: If I know the moments of a random var $X$ from 0 to some $k_0 \in \mathbb{N}$, do I know the distribution of $X$? No

Question 2: If I know all the moments of $X$ (assuming they are all finite), do I know the distribution of $X$? Not always!

Example: Assume $X$ is bounded (i.e., $\exists c > 0$ s.t. $|X(\omega)| \leq c$)

In this case, $\mathbb{E}(|X|^k) < +\infty \ \forall k \geq 0$ 

And moreover, $\mathbb{E}(|X|^k) \leq C^k \ \forall k \geq 0$ and the sequence $(m_k, k \geq 0)$ determines uniquely the dist. of $X$. 
Example: consider \( Z \sim N(0,1) \) and \( X = e^Z \)

( the distribution of \( X \) is the log-normal dist. )

\[
m_k = \mathbb{E}(x^k) = \mathbb{E}(e^{kZ}) = \ldots = \exp(k^2/2)
\]

In this, there exists another distribution with exactly the same sequence of moments!

Theorem (carleman)

If \( X \) is a r.v. whose all moments are finite and satisfy

\[
\sum_{k \geq 1} \frac{m_{2k}}{2k} = +\infty \quad (\star)
\]

then the distribution of \( X \) is uniquely determined by its sequence of moments.
is roughly equivalent to: \[ |m_k| \leq \exp(c_k \log k) \]
with \( c \leq 1 \)

**Examples:**

- \( X \) is a bold r.v.: \[ |m_k| \leq C^k \implies m_{2k} \leq C^{2k} \]
  
  so \[ m_{2k} - \frac{1}{2k} \geq \frac{1}{C} \implies \sum_{k \geq 1} m_{2k} - \frac{1}{2k} \geq \sum_{k \geq 1} \frac{1}{C} = +\infty \]

- \( X \) log-normal: \[ m_k = \exp(k^{2/2}) \implies m_{2k} = \exp(2k^2) \]
  
  so \[ m_{2k} - \frac{1}{2k} = \exp(-k) \implies \sum_{k \geq 1} m_{2k} - \frac{1}{2k} = \sum_{k \geq 1} e^{-k} < +\infty \]

*Note:* growth of \( m_k \) as \( k \to \infty \)

\( \leftrightarrow \) weight of the tail of the distribution
Remark: Condition \( \bullet \) only involves \( M_{2k}, k \geq 1 \).

\[
|m_{2k+1}| = |E(X_k X_{k+1})| = |E(X_k^2)|
\]

\[
\leq \sqrt{E(X_k^2) \cdot E(X_{k+2}^2)} = \sqrt{m_{2k} \cdot m_{2k+2}}
\]

**Theorem**

Let \( (X_n, n \geq 1) \) be a sequence of random variables such that \( m_k^{(n)} = E(X_n^k) \xrightarrow{n \to \infty} m_k \quad \forall k \geq 0 \)

where \( (m_k, k \geq 0) \) is a sequence of numbers satisfying Condition \( \bullet \). Then \( X_n \xrightarrow{d} X \) whose moments are \( (m_k, k \geq 0) \).
Concentration inequalities

Reminder: If \((X_n, n \geq 1)\) is a sequence of iid & integrable random variables and \(S_n = X_1 + \ldots + X_n\), then \(\forall \varepsilon > 0:\)
\[
P\left(\left| \frac{S_n}{n} - \mathbb{E}(X_n) \right| \geq \varepsilon \right) \xrightarrow{n \to \infty} 0
\]

Question: How fast does this probability tend to 0 as \(n\) gets large?

- Hoeffding's inequality (today)
- Cramér's theorem (next week)
Hoeffding's inequality

Let \((x_n, n \geq 1)\) be a sequence of iid and integrable r.v.'s all defined on a common probability space \((\Omega, \mathcal{F}, P)\) and such that \(|x_n(\omega) - E(x_n)| \leq 1\) \(\forall \omega \in \Omega \) \& \(n \geq 1\).

Let \(s_n = x_1 + \ldots + x_n, n \geq 1\). Then \(\lim_{n \to \infty}
\[
P\left( \exists \frac{s_n}{n} - E(x_1) \geq t \right) \leq 2 \cdot \exp\left(-\frac{nt^2}{2}\right)\]

Consequence: strong law of large numbers!

\[\sum_{n \geq 1} P\left( \left\{ \frac{s_n}{n} - E(x_1) \geq t \right\} \right) \leq 2 \sum_{n \geq 1} \exp\left(-\frac{nt^2}{2}\right) < +\infty\]

So by the Borel-Cantelli Lemma, \(\frac{s_n}{n} \to E(x_1)\) a.s.
Remark: take \( t = \frac{u}{\sqrt{n}} \)

\[ P \left( \left| \frac{\sum x_i}{n} - E(x_i) \right| \geq \frac{u}{\sqrt{n}} \right) \leq 2 \exp \left( -\frac{u^2}{2} \right) \]

Close to the CLT: \( \sqrt{n} \left( \frac{\sum x_i}{n} - E(x_i) \right) \xrightarrow{d} Z \sim N(0,1) \)

Proof of Hoeffding's inequality:

Lemma: Let \( Z \) be a r.v. such that \( |Z(w)| \leq 1 \) \( \forall w \in S \) & \( \mathbb{E}(Z) = 0 \). Then

\[ \mathbb{E} \left( e^{st} \right) \leq \exp \left( \frac{s^2}{2} \right) \quad \forall s \in \mathbb{R} \]
Proof: the mapping $z \mapsto e^{sz}$ is convex if $s \in \mathbb{R}$

So $e^{sz} \leq \frac{e^s + e^{-s}}{2} + z \frac{e^s - e^{-s}}{2}$

$-1 \leq z \leq 1 = \cosh(s) + z \sinh(s)$

So $E(e^{sz}) \leq E(\cosh(s) + z \sinh(s)) = \cosh(s) + 0$

$\leq \exp\left(\frac{s^2}{2}\right)$ (analysis) \#(lemma)
Proof of Hoeffding's inequality

\[ P \left( \{ \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}(X) \geq t \} \right) = P \left( \{ \sum_{i=1}^{n} X_i - n \mathbb{E}(X) \geq nt \} \right) \]

\[ = P \left( \{ \sum_{i=1}^{n} X_i - \mathbb{E}(X) \geq nt \} \right) + P \left( \{ \sum_{i=1}^{n} X_i - \mathbb{E}(X) \leq -nt \} \right) \]

\[ \left\{ \begin{array}{l}
P \left( \{ \sum_{i=1}^{n} X_i - \mathbb{E}(X) \geq nt \} \right) \leq \frac{\mathbb{E}(e^{s(X_i - \mathbb{E}(X))})}{e^{snt}} \\
\text{Chebyshev: } \\
\psi(x) = e^{sx}, \ s > 0 \\
= e^{-snt} \cdot \mathbb{E}(e^{\frac{n}{n-1}(X_i - \mathbb{E}(X))}) = e^{-snt} \cdot \prod_{j=1}^{n} \mathbb{E}(e^{s(X_j - \mathbb{E}(X_j))}) \\
= e^{-snt} \cdot \mathbb{E}(e^{s(X_i - \mathbb{E}(X))})^n \leq e^{-snt} \left( e^{s\frac{t^2}{2}} \right)^n = e^{-n(st - \frac{t^2}{2})} \\
\end{array} \right. \]

*Z (cf. Lemma)*
\((t>0)\)

\[
\max_{s \geq 0} \left( st - \frac{s^2}{2} \right) = ? \quad \frac{\partial}{\partial s} \left( st - \frac{s^2}{2} \right) = t - s = 0, \text{ i.e. } S^* = t
\]

\[
= t^2 - \frac{t^2}{2} = \frac{t^2}{2}
\]

So

\[
P \left( \{ S_n - \mathbb{E}(S_n) \geq nt \} \right) \leq \exp \left( -n \cdot \max_{s \geq 0} \left( st - \frac{s^2}{2} \right) \right)
\]

\[
= \exp \left( -nt^2/2 \right)
\]

So finally,

\[
P \left( \{ \left| S_n - \mathbb{E}(S_n) \right| \geq nt \} \right) \leq 2 \cdot \exp \left( -nt^2/2 \right)
\]