1 Coupon collector's problem

Let $U_1, U_2, ...$ be i.i.d. random variables, each distributed uniformly on $\{1, 2, ..., n\}$. And let $|\{U_1, ..., U_k\}|$ be the number of distinct elements among the first k variables, and let $T_n := \inf\{k : |\{U_1, ..., U_k\}| = n\}$. Show that $\frac{T_n}{n \mathrm{log} n} \xrightarrow{\mathbb{P}} 1$.

Proof. We defined T_n to be the number of draws until we've seen all coupons. We can make this problem easier if we define some auxillary random variables t_i for $i \geq 1$ where each t_i is the number of draws it takes to see the i'th unique coupon once we've seen the $(i-1)$'th unique coupon. Then we have

$$T_n = t_1 + ... + t_n.$$

Let $\epsilon > 0$. By Chebyshev, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{T_n}{n\log n} - \mathbb{E}\left[\frac{T_n}{n\log n}\right]\right| > \epsilon\right) &\leq \frac{\mathrm{Var}\left(\frac{T_n}{n\log n}\right)}{\epsilon^2} \\
&= \frac{\mathrm{Var}(T_n)}{(n\log n\ \epsilon)^2} \\
&= \frac{\sum_{i=1}^n \mathrm{Var}(t_i)}{(n\log n\ \epsilon)^2} && \text{(the } t_i \text{ are pairwise independent).}
\end{aligned}$$

Also note that each t_i is a geometric random variable with success probability $p_i := \frac{n-i+1}{n}$. So we have a closed form for its variance: $\mathrm{Var}(t_i) = \frac{1-p_i}{p_i^2}$. So for our term in the numerator in the last line of the inequality above, we have

$$\begin{aligned}
\sum_{i=1}^n \mathrm{Var}(t_i) &= \sum_{i=1}^n \frac{1-p_i}{p_i^2} \\
&= \sum_{i=1}^n \frac{n(i-1)}{(n-i+1)^2} \\
&= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + ... + \frac{n-1}{1^2}\right) \\
&\leq n\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + ... + \frac{n-1}{1^2}\right) && \text{(adding } \frac{1}{n^2} \text{ to the sum)} \\
&\leq n^2\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} + ... + 1\right) && \text{(making numerators all equal to } n\text{, then pulling it out)} \\
&= n^2 \sum_{i=1}^n \frac{1}{i^2} \\
&< n^2 \frac{\pi^2}{6} && \text{(Basel problem).}
\end{aligned}$$

Returning to our original inequality, letting $c > \frac{\pi^2}{6}$ be a constant, we have

$$\mathbb{P}\left(\left|\frac{T_n}{n\log n} - \mathbb{E}\left[\frac{T_n}{n\log n}\right]\right| > \epsilon\right) \leq \frac{n^2 c}{n^2(\log n\ \epsilon)^2}$$

for every $\epsilon > 0$. Taking limits as $n \longrightarrow \infty$, we have

$$\lim_{n\to\infty} \frac{c}{(\log n\ \epsilon)^2} = 0$$

and thus

$$\frac{T_n}{n\log n} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{T_n}{n\log n}\right].$$

So all we need to show is

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{T_n}{n\log n}\right] = 1.$$

To do so, note that

$$\mathbb{E}[T_n] = \mathbb{E}[t_1 + ... + t_n] = \sum_{i=1}^{n} \frac{1}{p_i} = n \sum_{i=1}^{n} \frac{1}{i}.$$

We use the handy inequality that

$$\log(n+1) \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log n + 1,$$

which gives us

$$\frac{n \log(n+1)}{n \log n} \leq \mathbb{E}\left[\frac{T_n}{n \log n}\right] \leq \frac{n \log n + 1}{n \log n}$$

and clearly both outer expressions tend to 1 as $n \longrightarrow \infty$. □

2 "Almost" law of iterated algorithm

Let $X_1, X_2, ...$ be standard normal i.i.d random variables and let $S_n := X_1 + ... + X_n$.

(i) Show that

$$\frac{1}{2\pi}\left[\frac{1}{x} - \frac{1}{x^3}\right] e^{-x^2/2} \leq \mathbb{P}(X_1 \geq x) \leq \frac{1}{2\pi}\frac{1}{x} e^{-x^2/2}.$$

Proof. Since X_1 is standard normal, we know

$$\mathbb{P}(X_1 \geq x) = \int_x^\infty \frac{1}{2\pi} e^{-y^2/2} \, dy.$$

We can simplify the problem if we make a change of variables corresponding to a shift of x. Once we change variables, we shift the point x to the origin. We let $z = y - x$, and therefore the integral becomes

$$\begin{aligned} &= \int_0^\infty \frac{1}{2\pi} e^{-(z+x)^2/2} \, dz \\ &= \frac{1}{2\pi} e^{-x^2/2} \left(\int_0^\infty e^{-(z^2+2zx)/2} \, dz \right). \end{aligned}$$

Now we deal with the integral on the right. Note that $z^2 + 2zx \geq 2zx$, and therefore $e^{-(z^2+2zx)} \leq e^{-2zx}$, and therefore

$$\int_0^\infty e^{-(z^2+2zx)/2} \, dz \leq \int_0^\infty e^{-zx} \, dz = \frac{1}{x}.$$

So we've shown the right side of the inequality. Now we show the left side. First split up the exponentials, and then make use of the fact that $1 - x \leq e^{-x}$, so we have

$$\begin{aligned} \int_0^\infty e^{-(z^2+2zx)/2} \, dz &= \int_0^\infty e^{-z^2/2} e^{-zx} \, dz \\ &\leq \int_0^\infty \left(1 - \frac{z^2}{2}\right) e^{-zx} \, dz. \end{aligned}$$

Integrating by parts yields the desired inequality

$$\frac{1}{2\pi}\left[\frac{1}{x} - \frac{1}{x^3}\right] e^{-x^2/2} \leq \mathbb{P}(X_1 \geq x) \leq \frac{1}{2\pi}\frac{1}{x} e^{-x^2/2}.$$

□

(ii) Show that $\limsup_{n\to\infty} \frac{X_n}{\sqrt{2 \log n}} = 1$ almost surely.

Proof. We use Borel-Cantelli and the inequality we've shown above to squeeze the lim sup to 1. Specifically, for every $\epsilon > 0$, we claim

$$\mathbb{P}\left(1-\epsilon \leq \limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \leq 1+\epsilon\right) = 1.$$

Let $\epsilon > 0$. Define the event $A_n^\epsilon := \{\frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\}$. We have

$$\begin{aligned}\mathbb{P}\left(\frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\right) &= \mathbb{P}(X_n \geq \sqrt{2\log n}(1+\epsilon)) \\ &\leq \frac{1}{2\pi}\frac{1}{\sqrt{2\log n}(1+\epsilon)} e^{-(\sqrt{2\log n}(1+\epsilon))^2/2} \\ &= \frac{1}{2\pi}\frac{1}{\sqrt{2\log n}(1+\epsilon)}\frac{1}{n^{(1+\epsilon)^2}}.\end{aligned}$$

So

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n^\epsilon) < \infty$$

by the p-series test, and by Borel-Cantelli I we have,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\right) = 0.$$

We follow the same approach to show the other bound. Let $B_n^\epsilon := \left\{\frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right\}$. So

$$\begin{aligned}\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right) &\geq \frac{1}{2\pi}\left[\frac{1}{\sqrt{2\log n}(1-\epsilon)} - \frac{1}{(\sqrt{2\log n}(1-\epsilon))^3}\right] e^{(\sqrt{2\log n}(1-\epsilon))^2/2} \\ &= \frac{1}{2\pi}\left[\frac{1}{\sqrt{2\log n}(1-\epsilon)} - \frac{1}{(\sqrt{2\log n}(1-\epsilon))^3}\right]\frac{1}{n^{(1-\epsilon)^2}}\end{aligned}$$

which is not summable i.e. $\sum_{n=1}^{\infty} \mathbb{P}(B_n^\epsilon) = \infty$. Thus, by Borel-Cantelli II,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right) = 1.$$

Putting everything together, we've shown

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} = 1\right) = 1.$$

□

(iii) Let C be a constant such that $C > \sqrt{2}$. Show that $\limsup_{n\to\infty}\left(\frac{S_n}{\sqrt{2\log n}}\right) < C$ almost surely.

Proof. We use the same approach as before. Let $\epsilon > 0$, and let $C = \sqrt{2}+\epsilon$. Note that $\frac{S_n}{\sqrt{n}}$ is standard normal. Let $C_n^\epsilon := \left\{\frac{S_n}{\sqrt{n}} \geq (\sqrt{2}+\epsilon)\sqrt{\log n}\right\}$. Then for every n, we have

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{\sqrt{n}} > (\sqrt{2}+\epsilon)\sqrt{\log n}\right) &\leq \frac{1}{2\pi}\frac{1}{(\sqrt{2}+\epsilon)\sqrt{\log n}} e^{-((\sqrt{2}+\epsilon)\sqrt{\log n})^2/2} \\ &= \frac{1}{2\pi}\frac{1}{(\sqrt{2}+\epsilon)\sqrt{\log n}}\frac{1}{n^{(\sqrt{2}+\epsilon)^2/2}} \\ &\leq \frac{1}{n^\alpha}\end{aligned}$$

for some $\alpha > 1$. Thus, $\sum_{n=1}^{\infty} \mathbb{P}(C_n^\epsilon) < \infty$ and

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} \geq (\sqrt{2}+\epsilon)\sqrt{\log n}\right) = 0.$$

Equivalently,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{n\log n}} < C\right) = 1.$$

□

3 Poisson approximation to the binomial distribution

Let $\{p_n\}_{n\geq 1}$ be a positive sequence such that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$ where $\lambda \in (0,\infty)$. Show that $\mathrm{Bin}(n, p_n)$ converges in distribution to $Poi(\lambda)$ as $n \longrightarrow \infty$.

Proof. Let $\lambda_n := np_n$ i.e. the expected value of $X_n \sim \mathrm{Bin}(n, p_n)$. The distribution for X_n is as follows:

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{n!}{(n-k)!k!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{(n)(n-1)(n-2)\cdots(n-k+1)}{k!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\
&= \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)\left(\frac{\lambda_n^k}{k!}\right)\left(1 - \frac{\lambda_n}{n}\right)^{n-k}
\end{aligned}$$

where all we did going from the third to fourth line was swap the position of $k!$ and n^k, and expanding n^k into k number of terms. Taking the limits as $n \longrightarrow \infty$, we see that

$$\begin{aligned}
\lim_{n\to\infty} \frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n}\frac{\lambda_n^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^{n-k} &= 1 \cdot \left[\lim_{n\to\infty}\frac{\lambda_n^k}{k!}\right] \cdot \left[\lim_{n\to\infty}\left(1-\frac{\lambda_n}{k!}\right)^{n-k}\right] \\
&= \left[\lim_{n\to\infty}\frac{\lambda_n^k}{k!}\right] \cdot \left[\lim_{n\to\infty}\left(1-\frac{\lambda_n}{k!}\right)^{n-k}\right] \\
&= \frac{\lambda^k}{k!} \cdot \left[\lim_{n\to\infty}\left(1+\frac{\lambda_n}{n}\right)^{n}\right] \cdot \left[\lim_{n\to\infty}\left(1+\frac{\lambda_n}{n}\right)^{-k}\right] \\
&= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} e^{-\lambda_n} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

which is the Poisson probability mass at k. All we did in the first line was realize that $\lim_{n\to\infty} \frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n} = 1$. $\square$

4 Exponential approximation to geometric distribution

Let X_p be a random variable with geometric distribution with parameter $p \in (0,1)$. Show that pX_p converges in distribution to Z where $Z \sim \mathrm{Exp}(1)$.
First we prove a useful lemma.
Lemma. If $c_n \longrightarrow 0$ and $a_n \longrightarrow \infty$ but $a_n c_n \longrightarrow \lambda$, then $(1+c_n)^{a_n} \longrightarrow e^{\lambda}$.

Proof. We prove that $\lim_{n\to\infty} a_n \log(1+c_n) = \lambda$, so that in the end we have

$$\lim_{n\to\infty}(1+c_n)^{a_n} = \lim_{n\to\infty} e^{\log[(1+c_n)^{a_n}]} = \lim_{n\to\infty} e^{a_n \log(1+c_n)} = e^{\lambda}.$$

First we derive the Taylor expansion of $\log(1+x)$. We use the integral definition of log (and make a simple substitution), to get

$$\log(1+x) = \int_0^x \frac{1}{1+t}\, dt.$$

Now we realize we can write the integrand as a geometric sum, yielding

$$\frac{1}{1+t} = \frac{1}{1-(-t)} = \sum_{n=0}^{\infty}(-t)^n = 1 - t + t^2 - t^3 + ...$$

Thus we can integrate the infinite series on the right term by term which yields

$$\log(1+x) = \int_0^x 1 \, dt - \int_0^x t \, dt + \int_0^x t^2 \, dt - ... = x - \frac{x^2}{2} + \frac{x^3}{3} - ...$$
$$= \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}.$$

Now we can determine that $a_n \log(1 + c_n) = \lambda$ using the Taylor expansion above. We have

$$\begin{aligned}
\lim_{n\to\infty} a_n \log(1+c_n) &= \lim_{n\to\infty} \left[a_n \left(c_n - \frac{c_n^2}{2} + \mathcal{O}(c_n^3) \right) \right] \\
&= \lim_{n\to\infty} \left[a_n c_n - \frac{a_n c_n^2}{2} + a_n \mathcal{O}(c_n^3) \right] \\
&= \lambda - \lim_{n\to\infty} \left[a_n c_n \frac{c_n}{2} + a_n c_n \mathcal{O}(c_n^2) \right] \\
&= \lambda \qquad\qquad (\text{since } c_n \longrightarrow 0).
\end{aligned}$$

We've proven the lemma. Now, showing weak convergence is straight forward. We have

$$\mathbb{P}(pX_p > x) = \mathbb{P}(X_p > \frac{x}{p}) = (1-p)^{\lfloor \frac{x}{p} \rfloor}.$$

Note that for small p, the quantity $(1-p)^{\lfloor \frac{x}{p} \rfloor} \approx (1-p)^{\frac{x}{p}}$. Hence we have

$$\lim_{p\to 0} (1-p)^{\lfloor \frac{x}{p} \rfloor} = \lim_{p\to 0} \left((1-p)^{\frac{1}{p}} \right)^x = e^{-x}.$$

Thus

$$\lim_{p\to 0} \mathbb{P}(pX_p > x) = e^{-x} = \mathbb{P}(Z > x).$$

□

5 Weak LLN for weakly correlated random variables

Let $r : \mathbb{N} \to \mathbb{R}$ be a bounded function such that $r(k) \to 0$ as $k \to \infty$. Let $X_1, X_2, ...$ be identical but not necessarily independent random variables with mean zero and finite variance. Suppose that the covariances of the random variables satisfy $\text{Cov}(X_i, X_j) \le r(|i-j|)$ for every $i, j \ge 1$. Let $S_n := X_1 + ... + X_n$. Show that $\frac{S_n}{n} \xrightarrow{\mathbb{P}} 0$.

Proof. Fix $\epsilon > 0$. By Chebyshev, we have that

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{S_n}{n} - 0 \right| > \epsilon \right) &\le \frac{\text{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \\
&= \frac{\sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j)}{n^2 \epsilon^2}.
\end{aligned}$$

Since $r(|i-j|) \longrightarrow 0$ as $|i-j| \longrightarrow \infty$ there exists some $K \in \mathbb{N}$ such that $\text{Cov}(X_i, X_j) \le r(|i-j|) \le \delta$ for all i, j such that $|i-j| \ge K$. So we split the sum of the covariances into two sums. The first sum is over the indices where the distance between i and j is less than N, and the second sum is over i and j where the distance between them is greater than or equal to K. For the term in the numerator, we have that

$$\begin{aligned}
\sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j) &= \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^{i+K} \text{Cov}(X_i, X_j) + 2\sum_{i=1}^n \sum_{j=i+K+1}^n \text{Cov}(X_i, X_j) \\
&\le \sum_{i=1}^n r(0) + 2\sum_{i=1}^n \sum_{j=i+1}^{i+K} r(|i-j|) + 2\sum_{i=1}^n \sum_{j=i+K+1}^n r(|i-j|).
\end{aligned}$$

Now letting $M := \max\{r(k) : k \le K\}$, we have

$$\begin{aligned} &\le nr(0) + 2\sum_{k=1}^{K-1}(n-k)r(k) + 2\sum_{k=K}^{n-K}(n-k)r(k) \\ &\le nM + 2KMn + 2\delta n^2. \end{aligned}$$

Plugging this back into the original expression, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S_n}{n} - 0\right| > \epsilon\right) &\le \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \\ &\le \frac{nM + 2KMn + 2\delta n^2}{n^2\epsilon^2} \\ &= \frac{M}{n\epsilon^2} + \frac{2KM}{n\epsilon^2} + \frac{2\delta}{\epsilon^2}. \end{aligned}$$

Taking the limit as $n \to \infty$ we have

$$\lim_{n\to\infty} \frac{M}{n\epsilon^2} + \frac{2KM}{n\epsilon^2} + \frac{2\delta}{\epsilon^2} = \frac{2\delta}{\epsilon^2},$$

and since δ was arbitrary, we can take it to 0. Hence for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - 0\right| > \epsilon\right) = 0,$$

and equivalently,

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} 0.$$

□