Question 1: You toss a coin $10,000$ times. How many heads do you see?

Question 2: Coupon collector problem. Have N coupons and we need to collect them all. How many coupons do we need to sample to get all N?

Question 3: Largest common subsequence problem: have sequences $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ of iid Bern(1/2) random variables. What is the largest k such that there exist $i_1 < i_2 < \dots < i_k$ and $j_1 < j_2 < \dots, < j_k$ such that $X_{i_1} = Y_{j_1}, \dots, X_{i_k} = Y_{j_k}$?

Question 1: we have various possible answers:

- 5,000. Indeed if we let X_i be the indicator of the event that we see heads on the ith toss, the number of heads is $S = \sum_{i=1}^{10000} X_i$ and $\mathbb{E}S = 5000$. But $\mathbb{P}(S = 5000) = \binom{10000}{5000} 2^{-1000} \approx 0.008$.
- Weak Law of Large Numbers: let $(X_i)_{i \geq 1}$ be iid with finite expectation μ and finite second moments. Then for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \xrightarrow{n \to \infty} 0.$$

  Therefore for large enough n, the number of heads lies in $[n(1/2-\varepsilon), n(1/2+\varepsilon)]$ with high probability. The main problem is that this is an asymptotic result - we don't know how large n should be.
- Central Limit Theorem: let $(X_i)_{i \geq 1}$ be iid with finite mean μ and finite second moment $\sigma^2 + \mu^2$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} \mathcal{N}(0,1).$$

  Therefore $\sum_{i=1}^{n}(X_i - \mu)$ has deviations of the order $\sqrt{n}\sigma$. Suppose we pretend 10000 is big: then

$$\begin{aligned} S = \sum_{i=1}^{10000} X_i &\in [5000 - Q^{-1}(0.005)\sqrt{100}/2, 5000 + Q^{-1}(0.005)\sqrt{100}/2] \\ &\approx [5000 \pm 128] \end{aligned}$$

  with probability 0.99, where $Q(x) = \mathbb{P}(Z \geq x)$ for $Z \sim \mathcal{N}(0,1)$. However we have the same issue again - is $n = 10000$ large enough?

We can however give some non-asymptotic answers to Question 1:

Proposition (Chebyshev's inequality)**.** Let X be any random variable with mean μ and variance σ^2. Then

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}.$$

With this, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{10000} X_i - 5000\right| > t\right) \leq \frac{10000 \times \frac{1}{4}}{t^2} = \frac{2500}{t^2}.$$

So in particular, if $t = 500$ the RHS is 0.01. So we have $S \in [4500, 5500]$ with probability 0.99. However note that this is a weaker result than what the Central Limit Theorem gives.

Question 2: the number of samples S is equal to $\sum_{i=1}^{N} X_i$ where $X_i \sim \mathrm{Geo}(i/N)$. Thus $\mathbb{E}S = \sum_{i=1}^{N} \frac{N}{i} = N \sum_{i=1}^{N} \frac{1}{i} \approx N \log N$.

Question 3: we have a function $f(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ which gives the longest common subsequence. It turns out this function is “smooth” in a certain sense, for which we can use “Talagrand’s Principle”.