1 Coupon Collector

Problem description: Let’s say a certain cereal company is selling cereal boxes with one of $n$ different figures. How many cereal boxes do you need to buy in order to collect all the figures? Let $n$ be the number of figures. Question: how long does this take?

1.1 Expected number of draws

Idea: let $Z_i$ be the time until the next new item when $i$ items are unseen. Then $T = Z_n + \cdots + Z_1$, and by linearity of expectation we have $\mathbb{E}[T] = \sum_{i=1}^{n} \mathbb{E}[Z_i]$. Each $Z_i$ is the number of random draws with probability $i/n$, so $Z_i \sim \text{Geometric}(i/n)$, and $\mathbb{E}[Z_i] = n/i$.

$$\mathbb{E}[T] = n \sum_{i=1}^{n} 1/i = nH_n \leq n(\log n + 1).$$

1.2 Concentration bounds

Now we want to find concentration bounds. Is it likely that this process will take a lot of draws?

First try: Markov’s inequality gives $T \leq 3nH_n$ with probability $2/3$. For error probability $1/n$, we need $n^2H_n$ draws.

Next try: Chebyshev’s inequality: $\mathbb{P}[|T - nH_n| \geq t] \leq \sigma^2/t^2$. $\sigma^2 = \text{Var}(T) = \sum_{i=1}^{n} \text{Var}(Z_i)$ since the time we found one item does not influence how many more draws you need until the next item, so $Z_i$ are independent. From Wikipedia, we have $\text{Var}(Z_i) = n(n-i)/i^2$, so

$$\sigma^2 = n \sum_{i=1}^{n} \frac{n-i}{i^2} \leq n^2 \sum_{i=1}^{n} 1/i^2 \leq n^2 \pi^2/6.$$ 

So $\sigma = O(n)$, so $\mathbb{P}[|T - nH_n| \geq tn] \leq \pi^2/6t^2$.

Note: to get the tightest bound, make use of the fact that each $Z_i$ is geometric and thus subexponential, so $T = \sum Z_i$ is subgamma (i.e. $\mathbb{E}[e^{\lambda x}] \leq e^{\lambda^2 \sigma^2/2}$ for all $\lambda \leq B$ for some bound $B$), which somehow implies that the tail is exponential.

1.3 Alternative concentration bound

$$\mathbb{P}[(\text{element } i \text{ not seen by time } T)] = (1 - 1/n)^T,$$
so by union bound

\[ P[\text{any element not seen by time } T] \leq n(1 - 1/n)^T \approx ne^{-T/n}. \]

## 2 Balls and Bins

We throw \( n \) balls into \( n \) bins.

\( X_i := \# \text{ balls in bin } i \)

\( \mathbb{E}[X_i] \)? \( \mathbb{E}[\max X_i] \)? \( \mathbb{E}[\text{empty bins}] \)? Concentration?

### 2.1 Expectation of each \( X_i \)

We know \( \sum X_i = n \), so by linearity of expectation \( \mathbb{E}[X_i] = 1 \).

### 2.2 Concentration of \( \max X_i \)

Turns out it’s easier to look at concentration first than to derive expectation.

Let’s look at \( \mathbb{P}[X_i = k] = \binom{n}{k} \frac{1}{n^k} (1 - \frac{1}{n})^{n-k} \).

**Key property:**

\[ \left( \frac{n}{k} \right)^k \leq \binom{n}{k} \leq \left( \frac{en}{k} \right)^k. \]

The left side follows from the fact that

\[ \binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots1}, \]

and each individual fraction is less than \( n/k \). The right side follows from Stirling’s approximation.

Then \( \mathbb{P}[X_i = k] \leq (en/k)^k(1-1/n)^{n-k} \leq (e/k)^k \).

\( \mathbb{P}[X_i \geq k] = \sum_{j=k}^\infty (e/j)^j \leq 2(e/k)^k \) if \( k \geq 6 \).

Then by union bound

\[ \mathbb{P}[\max X_i \geq k] \leq n \mathbb{P}[X_i \geq k] \leq 2n(e/k)^k; \]

then for this to be less than a constant, we have \( \mathbb{P}[\max X_i \geq k] \leq 1/2 \) if \( k \geq O(n) \). Turns out if \( k = \Theta(\log n / \log \log n) \), we have \( k^k \geq \sqrt{\log n^k} = (\log n)^{\frac{1}{2}k} = (\log n)^{\Theta(\log \log n)} = n \).

### 2.3 Expectation of \( \max X_i \)

## 3 Negative association

A set of random variables \( x_1, ..., x_n \) is negatively associated (N.A.) if \( \forall \) disjoint subsets \( I, J \leq [n] \), and for all monotonically nondecreasing (a mirror argument holds for monotonically nonincreasing) \( f, g \)
\[ \mathbb{E}[f(X_i) \cdot g(X_j)] \leq \mathbb{E}[f(X_I)] \mathbb{E}[g(X_J)] \]

This means it concentrates at least as well as independent variables, and one variable tends to be smaller when another is bigger.

Proof:

### 3.1 Zero-one lemma

Claim: if \( x \in \{0, 1\} \) and \( \sum x_i = 1 \), then \( x \) is negatively associated.

Proof:

WLOG, \( f(0) = g(0) = 0 \), else add a constant factor. Hence, for all inputs, \( f(x_i), g(x_j) \geq 0 \).

\[ \mathbb{E}[f(X_I)g(X_J)] = 0 \leq \mathbb{E}[f(X_I)] \mathbb{E}[g(X_J)]. \]

because always either \( X_I = 0 \) or \( X_J = 0 \).

### 3.2 Composition rules

1. If have N.A. random variables and apply monotonically nondecreasing function, the application of the function creates a new N.A. set of random variables.

2. If \( X, Y \) are individually N.A. and independent, then \((X, Y)\) is N.A.

This relates back to balls in bins!

Take \( W_{i,j} = 1 \) iff ball \( i \) lands in bin \( j \). Then,

1. All \( W_{i,*} \) are negatively associated with each other, and

2. \( W_{*,j} \) is also negatively associated.

Even though \( Z_i \) isn’t independent, we can still use the Chernoff bound because the Chernoff bound is based on the moment-generating function which changes little for negative associativity.

**Independence:** \[ \mathbb{E}(e^{\lambda(\sum z_i - \mu_i)}) = \prod_i \mathbb{E}(e^{\lambda(z_i - \mu_i)}) \]

**Negative associativity:** \[ \mathbb{E}(e^{\lambda(\sum z_i - \mu_i)}) \leq \prod_i \mathbb{E}(e^{\lambda(z_i - \mu_i)}) \]