Week 10.2: Coupons, X-rays, and Other Thoughts

The Coupon Collector Problem

We[1] describe the statement of the *coupon collector problem* as follows. Let n be a positive integer. Suppose that in each cereal box there is one of n different toys, which you would like to collect them all. Assume that each box is independent, and in each box each type of toy is equally likely. You keep buying these cereal boxes until you have collected all n types of toy. What is the expected number of cereal boxes you need to buy?

For each $i = 1, 2, \dots, n$, we let T_i denote the number of boxes you buy until you've collected i different toys. For example, $T_1 = 1$ with probability 1, because once you buy the first box, you obtain one type of toy. Now suppose that you've collected i different types of toy. The number of boxes you need to buy in order to obtain a new type of toys is a $\mathrm{Geom}((n-i)/n)$ random variable: there is a probability of i/n of getting a type you have already collected, and a probability of $(n-i)/n$ of getting a new type, in a box. This means

$$T_{i+1} - T_i \sim \mathrm{Geom}((n-i)/n).$$

The number of boxes you need to buy until you have collected all n types of toy is T_n. Note that

$$T_n = (T_n - T_{n-1}) + (T_{n-1} - T_{n-2}) + \cdots + (T_2 - T_1) + T_1.$$

Let us denote $G_1 := T_2 - T_1$, $G_2 := T_3 - T_2$, ..., and $G_{n-1} = T_n - T_{n-1}$. So

$$T_n = 1 + G_1 + G_2 + \cdots + G_{n-1}.$$

Observe that $G_1, G_2, \dots, G_{n-1}$ are independent geometric random variables with $G_i \sim \mathrm{Geom}((n-i)/n)$.

Recall from the first(!) week that a $\mathrm{Geom}(p)$ random variable has mean $1/p$ and variance $(1-p)/p^2$. From this, it is quite easy to compute the mean and the variance of T_n. We have

$$\mathbb{E}(T_n) = 1 + \sum_{i=1}^{n-1} \frac{n}{n-i} = n \cdot H_n, \tag{1}$$

and

$$\mathrm{Var}(T_n) = \sum_{i=1}^{n-1} \frac{in}{(n-i)^2} = \sum_{i=1}^{n-1} \frac{(n-i)n}{i^2}, \tag{2}$$

[1] Lecture Notes for ICMA 393/487, Trimester 1 Academic Year 2024-2025, Mahidol University International College. **Instructors:** Thotsaporn Thanatipanonda and Pakawut Jiradilok. (Notes Version: Nov 21, 2024.)

where H_n is the n^{th} harmonic number:

$$H_n = \frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}.$$

Let us invoke the following fact about the harmonic numbers: we have

$$H_n = \log n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right), \tag{3}$$

for positive integers n, where $\gamma \approx 0.577$ denotes the *Euler–Mascheroni constant*.

From Equation (2), we find

$$\operatorname{Var}(T_n) = 1 + n^2 \cdot S_n - n \cdot H_n,$$

where S_n denotes the summation

$$S_n := \sum_{i=1}^{n-1} \frac{1}{i^2}.$$

Such a sum can be approximated using the Euler–Maclaurin summation formula. Or, if we are lazy, we can do the following trick: note that

$$\frac{1}{i^2} = \frac{1}{i(i+1)} + \frac{1}{i(i+1)(i+2)} + \frac{2}{i^2(i+1)(i+2)}.$$

Therefore,

$$\begin{aligned}\sum_{i=n}^{\infty} \frac{1}{i^2} &= \frac{1}{n} + \frac{1}{2n(n+1)} + O\left(\frac{1}{n^3}\right)\\ &= \frac{1}{n} + \frac{1}{2n^2} + O\left(\frac{1}{n^3}\right).\end{aligned}$$

This implies

$$S_n = \frac{\pi^2}{6} - \frac{1}{n} - \frac{1}{2n^2} + O\left(\frac{1}{n^3}\right). \tag{4}$$

Using (3) and (4) in the formula for the variance of T_n above, we find

$$\operatorname{Var}(T_n) = \frac{\pi^2}{6} \cdot n^2 - n \cdot \log n - (\gamma + 1) \cdot n + O\left(\frac{1}{n}\right).$$

Now use (3) in (1) to obtain

$$\mathbb{E}(T_n) = n \cdot \log n + \gamma \cdot n + \frac{1}{2} + O\left(\frac{1}{n}\right).$$

Chebyshev's inequality states with if a real-valued random variable X has finite mean μ and finite variance σ^2 (with $\sigma > 0$), then for every real number $t > 0$, we have

$$\mathbb{P}\left\{\frac{|X-\mu|}{\sigma} \geq t\right\} \leq \frac{1}{t^2}.$$

From Chebyshev's inequality, we see that there is a *concentration* behavior for T_n. For instance, we have the following proposition.

Proposition 1. *There exist absolute constants $C_1, C_2 > 0$ such that the inequality*

$$\mathbb{P}\left\{1 - \frac{1}{\sqrt{\log n}} \leq \frac{T_n}{n \log n} \leq 1 + \frac{1}{\sqrt{\log n}}\right\} \leq 1 - \frac{C_2}{\log n}$$

holds for all positive integers $n \geq C_1$.

The above proposition implies, informally,

$$\mathbb{P}\left\{1 - o(1) \leq \frac{T_n}{n \log n} \leq 1 + o(1)\right\} = 1 - o(1),$$

as $n \to \infty$.

X-rays of permutations

A reference for this subsection is Bebeacua–Mansour–Postnikov–Severini [BMPS05].

Let n be a positive integer. For each permutation $w \in S_n$, the **permutation matrix** of w, denoted P_w, is the $n \times n$ matrix whose (i,j)-entry is

$$(P_w)_{i,j} = \begin{cases} 1, & \text{if } w(i) = j, \\ 0, & \text{if } w(i) \neq j. \end{cases}$$

For example, if

$$w = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \in S_3,$$

then

$$P_w = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3}.$$

For each matrix $M \in \mathbb{R}^{n\times n}$, we define the **X-ray** of M, denoted $\mathfrak{X}(M)$, to be the $(2n-1)$-tuple whose k^{th} entry is

$$(\mathfrak{X}(M))_k = \sum_{\substack{i,j\in[n] \\ i+j=k+1}} M_{ij}.$$

For each permutation $w \in S_n$, the **X-ray** $\mathfrak{X}(w)$ of w is defined to be the X-ray of P_w. For example, if

$$w = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix} \in S_4,$$

then

$$P_w = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4},$$

and the X-ray is

$$\mathfrak{X}(w) = \mathfrak{X}(P_w) = (0,0,2,1,0,1,0) \in \mathbb{R}^7.$$

The following proposition follows immediately from the definition of X-rays.

Proposition 2. *Let n, k be positive integers such that $1 \leq k \leq 2n-1$. Then for any $w \in S_n$, we have*

$$0 \leq (\mathfrak{X}(w))_k \leq \min\{k, 2n-k\}.$$

From now on, let us consider when $k \leq n$. (Note that the case $k \geq n+1$ is analogous.) In this case, we have

$$0 \leq (\mathfrak{X}(w))_k \leq k.$$

We ask: how does the X-ray of a random permutation behave?

Take $\mathfrak{w} \sim \mathrm{Unif}(S_n)$. Note that

$$(\mathfrak{X}(\mathfrak{w}))_k = I_1 + I_2 + \cdots + I_k,$$

where $I_1, I_2, \ldots, I_k$ are indicator random variables given by

$$I_\ell = \begin{cases} 1, & \text{if } \mathfrak{w}(\ell) = k+1-\ell, \\ 0, & \text{if } \mathfrak{w}(\ell) \neq k+1-\ell. \end{cases}$$

We have the following proposition.

Proposition 3. *For $1 \leq i < j \leq k$, we have*

$$\mathbb{E}[I_i] = \frac{1}{n} \qquad \textit{and} \qquad \mathbb{E}[I_i I_j] = \frac{1}{n(n-1)}.$$

Therefore,

$$\mathbb{E}((\mathfrak{X}(\mathfrak{w}))_k) = \frac{k}{n},$$

and

$$\mathrm{Var}((\mathfrak{X}(\mathfrak{w}))_k) = \frac{k}{n} + \frac{k(k-1)}{n(n-1)} - \frac{k^2}{n^2}.$$

Note that in the regime $n \to \infty$, this variance is quite large compared to the mean. So is there a concentration behavior?

Exercise 1. Consider the following regime. Fix a positive integer k, and let $n \to \infty$. Consider the random variable

$$Z := \frac{(\mathfrak{X}(\mathfrak{w}))_k}{\mathbb{E}((\mathfrak{X}(\mathfrak{w}))_k)} = \frac{(\mathfrak{X}(\mathfrak{w}))_k}{k/n} = \frac{n}{k} \cdot (\mathfrak{X}(\mathfrak{w}))_k$$

Perform a computer simulation. Fix a small k; e.g. $k = 5$. Simulate Z for different increasing values of n. (For each n, simulate many instances of Z.) As n grows bigger, does it seem that Z concentrate better at 1 or not?

Tying up loose ends

Random Walks on $\mathbb{Z}/n\mathbb{Z}$

Recall the following problem. If we perform a symmetric random walk on $\mathbb{Z}/n\mathbb{Z}$ starting at 0, what is the distribution of the last number we visit. We showed an inductive argument showing that the distribution is uniform on $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$.

Here is an elegant and much simpler argument presented in the book of Sheldon Ross' [Ros10]. Fix any $i \neq 0$ in $\mathbb{Z}/n\mathbb{Z}$. Look at the first time the walk hits $i - 1$ or $i + 1$. Suppose for now that at the time the walk hits $i + 1$. (The case when it hits $i - 1$ is similar.) The probability that i is the last label hit is the same as the probability that starting at $i + 1$ the walk hits $i - 1$ before it hits i. By gambler's ruin, this probability is $1/(n - 1)$. We have finished.

Variance of sample mean

Suppose that X_{ij} are i.i.d. random variables following some distribution $\mathcal{D}$ with mean μ and variance σ^2. We compute sample means

$$Y_i := \frac{X_{i1} + X_{i2} + \cdots + X_{in}}{n},$$

for $i = 1, 2, \ldots, N$. Suppose we would like to estimate the *variance of the sample mean* and we have the data of

$$Y_1, Y_2, \ldots, Y_N.$$

Should we use the *sample variance* (dividing by $N - 1$) or the *population variance* (dividing by N)?

In the situation, the following estimator

$$\frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2,$$

where

$$\overline{Y} := \frac{Y_1 + Y_2 + \cdots + Y_N}{N},$$

is an unbiased estimator for the variance of Y_i. Note that we divide by $N-1$, since we used $\overline{Y}$ instead of the actual expectation of Y_i.

Note that we can also ask about the variance of the sample variance as well.

Exercise 2. Let X_{ij} be i.i.d. Unif$(0,1)$. Define

$$V_i := \frac{1}{n-1} \sum_{j=1}^{n} \left(X_{ij} - \overline{X_i} \right)^2,$$

where

$$\overline{X_i} := \frac{X_{i1} + X_{i2} + \cdots + X_{in}}{n},$$

for each $i = 1, 2, \ldots, N$.

What are the mean and the variance of V_i?

Finite first moment

Let X be a real-valued random variable. Are

$$|\mathbb{E}(X)| < \infty \tag{$\heartsuit$}$$

and

$$\mathbb{E}(|X|) < \infty \tag{$\spadesuit$}$$

equivalent?

It seems so, somehow. Recall that we may write

$$U := \max\{X, 0\} \qquad \text{and} \qquad V := -\min\{X, 0\},$$

so that U, V are nonnegative random variables, with $X = U - V$. Note that

$$|X| = U + V.$$

Both ($\heartsuit$) and ($\spadesuit$) are equivalent to the statement that both $\mathbb{E}(U)$ and $\mathbb{E}(V)$ are finite.

Central moments of the number of descents

Let X denote the number of descents in $\mathfrak{w} \sim \mathrm{Unif}(S_n)$. Recall that we showed

$$\mathbb{E}[X] = \frac{n-1}{2}.$$

Trivially, the first central moment is

$$\mathbb{E}\big[(X - \mathbb{E}[X])^1\big] = 0.$$

We showed that the second central moment (the variance) is

$$\mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \frac{1}{12} \cdot (n+1),$$

and that for every nonnegative integer k, we have

$$\mathbb{E}\big[(X - \mathbb{E}[X])^{2k+1}\big] = 0.$$

From our experiments, we have a "conjectural" formula:

$$\mathbb{E}\big[(X - \mathbb{E}[X])^4\big] = \frac{(n+1)(5n+3)}{240}.$$

Question 1. Is it true that for every positive integer k, the polynomial obtained from

$$\mathbb{E}\big[(X - \mathbb{E}[X])^k\big]$$

is divisible by $n+1$ (in the polynomial ring $\mathbb{Q}[n]$)?

References

[BMPS05] Cecilia Bebeacua, Toufik Mansour, Alex Postnikov, and Simone Severini. On the X-rays of permutations. In *Proceedings of the workshop on discrete tomography and its applictions, New York, NY, USA, June 13–15, 2005*, pages 193–203. Amsterdam: Elsevier, 2005.

[Ros10] Sheldon M. Ross. *Introduction to probability models.* Amsterdam: Elsevier/Academic Press, 10th ed. edition, 2010.