Chapter 2

Balls and Bins

2.1 Balls 'n' Bins – What is it?

Let's start with a game that will help us with hashing, our first data structure. We introduce the "balls and bins" game, in which we throw b balls equiprobably and independently into n bins. (Often, $b = n$.)

Applications. We can gain insight into hashing by studying the balls and bins game because hashing is modelled by randomly "throwing" data into hash table locations. Another application of balls and bins is in load-balancing, where bins can be thought of as servers, and balls as clients.

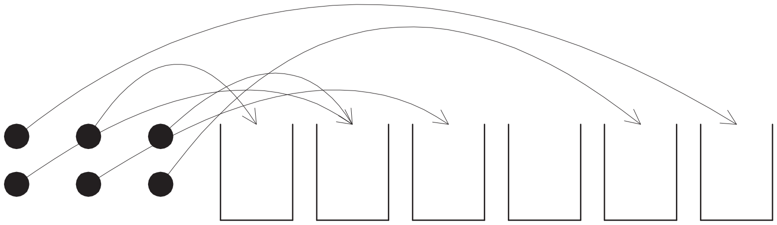

Figure 2.1: Balls and bins.

Questions to ask about this game. A number of interesting questions can be raised by playing this "balls and bins" game. What is the:

- expected number of balls in a bin
- expected number of balls in the fullest bin
- expected number of balls that need to be thrown before getting a collision (a bin with more than one ball)
- expected number of empty bins
- expected number of bins with a collision
- expected number of balls needed to fill all bins.
- and so on...

What happens when we replace "expected number" above with "with high probability'?'

In order to answer these questions, we need to review basic definitions and theorems of probability.

2.2 Review of Basic Probability

2.2.1 Sample Spaces and Events

Definition 1 (***Probability sample space*** $(\boldsymbol{S}, \boldsymbol{P})$)**.** *Let S be a set of outcomes, which is finite or countably infinite*

$$S = \{s_1, s_2, \dots\} .$$

Let probability function

$$P : S \to [0, 1] ,$$

where

$$\sum P(s_i) = 1 .$$

Definition 2. *An* ***event*** *is a subset of outcomes from the sample space (S, P).*

Probability is beautiful, but unintuitive. Here's a four-step process for solving many probability questions.

- Step 1: find the sample space
- Step 2: define events of interest.

- Step 3: determine outcome probabilities.
- Step 4: determine event probabilities.

2.2.2 Random Variables and Expectation

Definition 3. *A **random variable** is a function defined as*

$$f : S \to \Re^+ \ .$$

(Actually $\Re$ does not really need to be non-negative, but usually it is)

Note. A random variable isn't a variable. It's a function.

Definition 4. *Expected value $E[f]$ of random variable f*

$$E[f] = \sum p(s_i) f(s_i) \ .$$

Theorem 5. *Linearity of expectation*

$$E[f + g] = E[f] + E[g] \ .$$

Linear of expectation is a beautiful thing!

Example 1. We have n letters and n envelopes. Each letter has its envelope. We put letters randomly in envelopes. What is expected number of letters in the correct envelope?

Example 2. I flip a coin until I get tails. If I get i heads, I get 2^i dollars. What's the expect number of dollars that I earn?

2.2.3 Conditional Probability and Independence[1]

Very roughly, this section is how to think about $\Pr(A \cap B)$. (This idea helps me, but I don't know if it will help anyone else.)

Definition 6 (***Conditional probability***)**.** *The notation $\Pr(A|B)$ denotes the probability of event A happening, given that event B happens. Formally,*

$$\Pr(A|B) = \frac{Pr(A \cap B)}{\Pr(B)}.$$

If $\Pr(B) = 0$, then the conditional probability $\Pr(A|B)$ is undefined.

[1] Some of this material is from Eric Lehman et al's notes on the mathematics of computer science at MIT.

Definition 7 (***Independence***)**.** *Two events A and B are **independent** if and only if:* $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

Here is another way to think about independence:

Definition 8 (***Independence, alternative definition***)**.** *Two events A and B are **independent** if and only if* $\Pr(A|B) = \Pr(A)$ *or* $\Pr(B) = 0$.

Intuition about independence. Let's develop some intuition about independence.

Question. Suppose that we have two disjoint events; see Figure 2.2.3. Are these events independent?

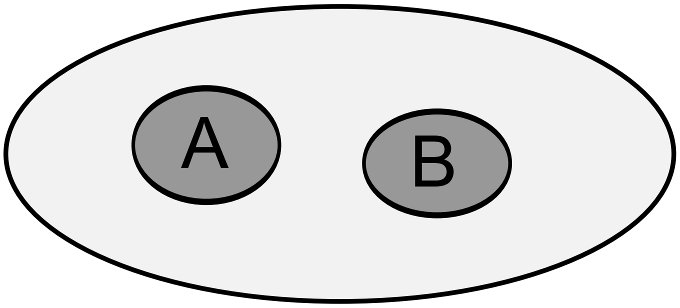

Figure 2.2: Illustration of two disjoint events.

Answer. No. Let's see why. We know that

$$\Pr(A \cap B) = 0$$

because the events are disjoint. On the other hand,

$$\Pr(A)\Pr(B) > 0$$

except in the degenerate case where one of the events has zero probability. Hence, disjointness and independence are very different concepts.

Figure 2.2.3 gives an intuitive picture of what independence means.

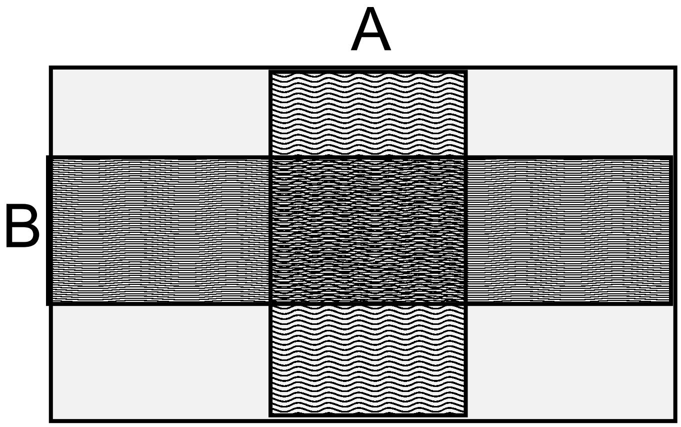

Figure 2.3: Illustration of two independent events.

2.2.4 Mutual Independence and Pairwise Independence

Definition 9. *Events $E_1, \ldots, E_n$ are **mutually independent** if and only if for every subset of the events, the probability of the intersection is the product of the probabilities of the individual events. In other words, all of the following equations must hold:*

$$
\begin{aligned}
\Pr(E_i \cap E_j) &= \Pr(E_i)\Pr(E_j) && \textit{f́or all distinct } i,\ j \\
\Pr(E_i \cap E_j \cap E_k) &= \Pr(E_i)\Pr(E_j)\Pr(E_k) && \textit{for all distinct } i,\ j,\ k \\
\Pr(E_i \cap E_j \cap E_k \cap E_\ell) &= \Pr(E_i)\Pr(E_j)\Pr(E_k)\Pr(E_\ell) && \textit{for all distinct } i,\ j,\ k,\ \ell \\
\Pr(E_1 \cap \ldots \cap E_n) &= \Pr(E_1)\ldots\Pr(E_n) && \textit{for all distinct } i, \ldots, n
\end{aligned}
$$

Example 3. Suppose that we flip three fair, mutually independent coins. Define the following events:

- A_1 is the event that coin 1 matches coin 2.

- A_2 is the event that coin 2 matches coin 3.

- A_3 is the event that coin 3 matches coin 1.

Are A_1, A_2, A_3 mutually independent?

Answer. No. But they are ***pairwise independent***.

Definition 10. *Events $E_1, \ldots, E_n$ are **pairwise independent** if and only if for every two events, the probability of the intersection is the product of the probabilities of the individual events. In other words:*

$$\Pr(E_i \cap E_j) \quad = \quad \Pr(E_i)\Pr(E_j) \qquad \textit{for all distinct } i,\ j$$

2.2.5 Independence of Random Variables

The notion of independence carries over from events to random variables.

Definition 11. *Random variables X_1 and X_2 are **independent** if*

$$\Pr(X_1 = a_1 \cap X_2 = a_2) = \Pr(X_1 = a_1)Pr(X_2 = a_2)$$

for all a_1 in the codomain (image) of X_1 and a_2 in the codomain (range) of X_2 .

The same notions of pairwise and mutual independence also carry over from events to random variables.

2.2.6 Inclusion/Exclusion

This section explains how to think about $\Pr(A \cup B)$.

Theorem 12 (Inclusion-Exclusion). *Given two events A_1 and A_2,*

$$\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2).$$

We can now generalize to three events.

Theorem 13 (Inclusion-Exclusion). *Given three events A_1, A_2, and A_3,*

$$\begin{aligned}\Pr(A_1 \cup A_2 \cup A_3) = & \quad \Pr(A_1) + \Pr(A_2) + \Pr(A_3) \\ & - \ \Pr(A_1 \cap A_2) - \Pr(A_1 \cap A_3) - \Pr(A_2 \cap A_3) \\ & + \ \Pr(A_1 \cap A_2 \cap A_3).\end{aligned}$$

We can similarly generalize to any number n of events, alternating plusses and minuses.

Inclusion exclusion is illustrated by Figure 2.2.6.

A correllary of the previous theorems is

Need a figure explaining how inclusion-exclusion works.

Figure 2.4: Illustration of inclusion exclusion.

Theorem 14 (Union bound)**.** *Given two events A_1 and A_2,*

$$\begin{aligned}\Pr(A_1 \cup A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\ &\leq \Pr(A_1) + \Pr(A_2).\end{aligned}$$

Given n events $A_1, A_2, \ldots, A_n$,

$$\Pr(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_n).$$

2.3 Answering the Questions

Now we start answering the questions that were raised. While we are answering the questions we will come across **Death Bed Formulae** which will be boxed separately throughout this document.

2.3.1 Expected Numbers of Balls in a Bin

Question 1: What is the **expected number** of balls in bin 1?

Theorem 15. *The expected number of balls in a bin is 1.*

First, define random variable

$$x_i = \begin{cases} 1 & \text{if ball } i \text{ lands in bin 1;} \\ 0 & \text{if ball } i \text{ lands in another bin.} \end{cases}$$

Define X, the number of balls in bin 1, as

$$X = x_1 + x_2 + \cdots + x_N \ .$$

The expected value of any ball i landing in bin 1 is

$$E[x_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n} \ .$$

By linearity of expectation, the expected number of balls in bin 1, $E[X]$, is

$$\begin{aligned}E[X] &= E[x_1] + E[x_2] + \cdots + E[x_N] \\ &= n \cdot E[x_1] = 1.\end{aligned}$$

2.3.2 Balls in the Fullest Bin

Question 2: What is the number of balls in the fullest bin **with high probability**, given that there are a total of n balls and n bins?

Before we answer this question we define the term "with high probability."

Definition 16. *Let E_n be an event on problem size n. We say that E_n occurs **with high probability** if $Pr(E_n) \geq 1 - \frac{1}{n^c}$, for some constant c.*

*Typically, E_n will be parametrized by some constant d. For example, E_n might be the event that a bin has $\Theta(\log n)$ balls, and the Θ hides a multiplicative constant. In this case, we can say even more strongly that **for every** c, there is a d so that $\Pr(E) \geq 1 - \frac{1}{n^c}$.*

This stronger definition is what we'll mean by "with high probability" unless otherwise noted.

Theorem 17. *The fullest bin has $O\left(\frac{\log n}{\log\log n}\right)$ balls with high probability.*

Proof: We start by giving the probability of the **1st** bin having ℓ balls. That is,

$$\Pr(\text{bin 1 has } \ell \text{ balls}) = \binom{n}{\ell}\left(\frac{1}{n}\right)^{\ell}\left(1-\frac{1}{n}\right)^{n-\ell}.$$

Now we give the probability that bin 1 has more than ℓ balls. And that is,

$$\Pr(\text{bin 1 has } \geq \ell \text{ balls}) \leq \binom{n}{\ell}\left(\frac{1}{n}\right)^{\ell}.$$

$$\boxed{\textbf{DON'T FORGET: } \left(\tfrac{y}{x}\right)^{x} \leq \tbinom{y}{x} \leq \left(\tfrac{ey}{x}\right)^{x}.}$$

From the above fact we get,

$$\begin{aligned}\Pr(\text{bin 1 has } \geq \ell \text{ balls}) &\leq \binom{n}{\ell}\left(\frac{1}{n}\right)^{\ell} \\ &\leq \left(\frac{en}{\ell n}\right)^{\ell} \\ &= \left(\frac{e}{\ell}\right)^{\ell}.\end{aligned}$$

$$\boxed{\textbf{DON'T FORGET: } P(A\cup B) = P(A)+P(B)-P(A\cap B) \leq P(A)+P(B).}$$

Let's say that $\ell = c \log n$. Then we get,

$$\begin{aligned}\Pr(\textbf{any} \text{ bin has } \geq c \log n \text{ balls}) &\leq n \left(\tfrac{e}{c \log n}\right)^{c \log n} \\ &\leq n \left(\tfrac{1}{2}\right)^{c \log n} \\ &\leq n \cdot n^{-c} \\ &= n^{1-c},\end{aligned}$$

which is polynomially small. Since this approximation is so loose, we can do better.

Now let us say the $\ell = c\frac{\log n}{\log \log n}$. Then:

$$\begin{aligned}\Pr(\textbf{any} \text{ bin has} \geq c\frac{\log n}{\log \log n} \text{ balls}) &\leq n \left(\frac{e \log \log n}{c \log n}\right)^{\frac{c \log n}{\log \log n}} \\ &\leq n 2^{\log\left(\frac{e \log \log n}{c \log n}\right)\frac{c \log n}{\log \log n}} \\ &\leq n 2^{\left(\frac{c \log n}{\log \log n}\right)(\log e + \log \log \log n - \log \log n - \log c)} \\ &\leq n \left(\frac{1}{2}\right)^{\left(\frac{c \log n}{\log \log n}\right)(\log \log n - O(\log \log \log n))} \\ &\leq n \left(\frac{1}{2}\right)^{c \log n - o(c \log n)} .\end{aligned}$$

For sufficiently large c and n, we obtain

$$\Pr(\textbf{any} \text{ bin has} \geq c\frac{\log n}{\log \log n} \text{ balls}) \leq n \left(\frac{1}{2}\right)^{(c-1) \log n} = n^{2-c},$$

which is also polynomially small.

In fact, the previous bound is tight. We will not attempt to prove the lower bound right away.

Theorem 18. *The number of balls in the fullest bin is* $\Omega\left(\frac{\log n}{\log \log n}\right)$ *whp.*

If we combine Theorems 17 and 18, we know that:

Theorem 19. *The number of balls in the fullest bin is* $\Theta(\frac{\log n}{\log \log n})$ *whp.*

2.3.3 Balls Needed to Fill all n Bins

Question 3: What is the number of balls needed to fill all n bins w.h.p.?

Theorem 20. *The number of balls needed to fill all the bins is* $\Theta(n \log n)$ *with high probability.*

Proof: A naive lower bound for this problem is $\Omega(n)$.

We find the upper bound by finding the probability of bin 1 being empty after ℓ balls have been thrown,

$$\Pr(\text{bin 1 is empty after } \ell \text{ balls}) = \left(1 - \frac{1}{n}\right)^{\ell}.$$

$$\boxed{\textbf{DON'T FORGET:}\quad (1-\tfrac{1}{n})^n \leq \tfrac{1}{e}.}$$

$$\Pr(\text{bin 1 is empty after } \ell \text{ balls}) = \left(1 - \frac{1}{n}\right)^{n\frac{\ell}{n}} \leq \left(\frac{1}{e}\right)^{\frac{\ell}{n}}.$$

Let $\ell = cn\ln n$. Plugging in this value of ℓ in the above inequality we get

$$\Pr(\text{bin 1 is empty after } \ell \text{ balls}) \leq \left(\frac{1}{e}\right)^{c\ln n} = n^{-c}.$$

Therefore,

$$\Pr(\textbf{any} \text{ bin is empty}) \leq n \cdot n^{-c} = n^{1-c}.$$

So we see that the number of balls required to fill all the bins w.h.p. is $O(n \log n)$. In fact, the number of balls required to fill all bins w.h.p is also $\Omega(n \log n)$.

Question 4: What is the expected number of balls required to fill all the bins?[2]

Theorem 21. *The expected number of balls required to fill all the bins is nh_n, where $h_n \approx \ln n$ is the nth harmonic number.*[3]

$$\boxed{h_1 = 1, h_2 = 1 + \tfrac{1}{2}, h_3 = 1 + \tfrac{1}{2} + \tfrac{1}{3}, \ldots ....h_n = 1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} \ldots ... + \tfrac{1}{n-1} + \tfrac{1}{n}.}$$

Proof: Divide the execution into phases $n, n-1, n-2, \ldots, 1$, where in phase i there are i free bins. In phase i the probability that a ball falls in an empty bin is

$$\Pr(\text{a ball falls in an empty bin}) = \frac{i}{n}.$$

[2] This question is also known as the ***Coupon Collector's problem***

[3] Figure 2.5 illustrates why the above approximation is true

Let X_i be a random variable measuring the number of balls thrown in phase i.
Then

$$E[X_i] = \frac{n}{i}.$$

Let the random variable X be the number of balls needed to fill all bins, i.e,

$$X = X_1 + X_2 + X_3 + \cdots ... + X_n.$$

By linearity of expectation

$$\begin{aligned} E[X] &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= n \cdot (1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + \tfrac{1}{n}) \\ &\approx n \cdot \ln n. \end{aligned}$$

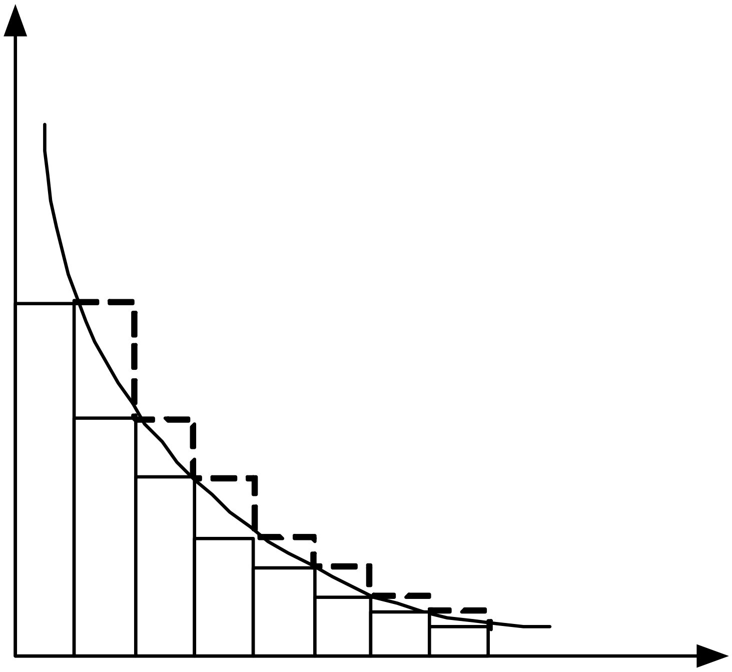

Figure 2.5: The curve of function $1/n$. The harmontic number $h_n \approx \ln n$ is the integral of this curve.

2.3.4 Number of Pairwise Collisions

Question 5: What is the expected number of pairwise collisions?

Theorem 22. *Suppose that we have n balls and cn^2 bins. Then the expected number of pairwise collisions is $\frac{1}{2c}$.*

Proof: Let there be a random variable X_{ij} such that,

$$X_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ collide;} \\ 0 & \text{otherwise.} \end{cases}$$

The total number of paiwise collisions is the sum of all the random variables for all $1 \leq i < j \leq n$.

$$X = \sum_{1 \leq i < j \leq n} X_{ij}.$$

The expectation of X_{ij} is given by,

$$\begin{aligned} E[X_{ij}] &= 1 \cdot \frac{1}{cn^2} + 0 \cdot \left(1 - \frac{1}{cn^2}\right) \\ &= \frac{1}{cn^2}. \end{aligned} \tag{2.1}$$

Thus, by linearity of expectation,

$$\begin{aligned} E[X] &= \left(\frac{n(n-1)}{2} \cdot \frac{1}{cn^2}\right) \\ &\approx \left(\frac{1}{2c}\right). \end{aligned}$$

We'll deal with other of the questions later. Some of the questions are more difficult because we do not have independence.