

## CHAPTER

# 9

# Counting and Chance

*Counting is the oldest mathematics there is. Doing it cleverly is still one of the hardest.*

### IN THIS CHAPTER

|      |                               |               |
|------|-------------------------------|---------------|
| 9.1  | Lines Cutting the Plane       | STEP II 1990  |
| 9.2  | Ramsey's Theorem in Miniature | STEP I 1989   |
| 9.3  | Colouring Fence Posts         | STEP II 1995  |
| 9.4  | Sequences as Functions        | STEP I 2016   |
| 9.5  | Making Change                 | STEP I 1997   |
| 9.6  | Strange Dice                  | STEP I 1993   |
| 9.7  | Equal Sums of Powers          | STEP I 1992   |
| 9.8  | Nobody Gets Their Own Shirt   | STEP II 2016  |
| 9.9  | The Birthday Problem          | STEP II 2007  |
| 9.10 | Always Able to Give Change    | STEP I 2011   |
| 9.11 | Collecting the Set            | STEP II 1994  |
| 9.12 | Random Subsets                | STEP III 2019 |
| 9.13 | Loaded Dice                   | STEP I 1999   |



# The Art of Counting

---

*Combinatorics is the slums of topology.*

— J. H. C. Whitehead, in the 1950s. It is not a view anyone holds today.

Everyone can count. But ask how many ways there are to make £1 from coins, or to seat ten people so that nobody sits in their own place, or to colour a map with four colours, and simple counting quickly breaks down. There are too many cases to list, and the cases overlap. *Combinatorics*, the mathematics of counting and arrangement, is the art of organising such problems so that they can be solved without listing.

For a long time, combinatorics was regarded as a collection of clever tricks rather than a serious subject, as Whitehead's remark suggests. That changed in the second half of the twentieth century, driven partly by computer science, where counting the steps an algorithm takes is a combinatorial problem, and partly by the discovery of deep connections with algebra, geometry and probability. Today it is one of the most active areas of mathematics. In 2022 June Huh was awarded a Fields Medal, the highest honour in mathematics, partly for proving a 1968 conjecture about the chromatic polynomials of Section 9.3, using ideas from algebraic geometry.

## Why counting matters

Probability, in its simplest form, is counting: when all outcomes are equally likely, the probability of an event is the number of favourable outcomes divided by the total. So every question about cards, dice and lotteries is a counting problem, and the second half of this chapter is drawn from the probability papers for exactly this reason. But counting also matters well beyond games. Computer scientists count the operations an algorithm needs, and the ways it might fail. Chemists count the possible structures of molecules. Biologists count the ways a strand of DNA can fold. Cryptographers count the keys an attacker must try, and, as Section 9.9 shows, how many are needed before two of something coincide.

## What you will meet at university

Most universities offer a first- or second-year course in combinatorics or discrete mathematics, and its themes are the themes of this chapter. You will learn the basic counting principles (the product rule, bijections, inclusion–exclusion, the pigeonhole principle), and how to set up and solve recurrence relations. You will meet *generating functions*, which turn sequences into functions and counting into algebra. You will study graphs, from colourings to Ramsey theory, and partially ordered sets such as the lattice of subsets. And probability courses will use the same tools, together with linearity of expectation, to analyse random structures. Further on, combinatorics

connects with algebra (symmetric functions, representation theory), with number theory (partitions, additive combinatorics) and with computer science (the analysis of algorithms, complexity theory).

The questions that run through this chapter

Three big questions

- 1. **How do you count without listing?** What structures (recurrences, bijections, generating functions) let you count huge collections exactly?
- 2. **When must something happen?** When does size alone force a pattern, a coincidence or a match?
- 3. **How do counting and probability help each other?** When is a probability best found by counting, and when is a count best found by probability?

The chapter has two halves. The first seven sections come from the pure papers. They begin with recurrences (lines cutting the plane, colouring fence posts) and the pigeonhole principle (Ramsey’s theorem), then develop generating functions and use them to count coins, design strange dice and split numbers into teams with equal power sums. The last six sections are classic counting problems from the probability papers: derangements and the appearance of  $1/e$ , the birthday problem, the ballot problem and its reflection principle, the coupon collector, random subsets, and the impossibility of loading dice so that every total is equally likely.

Ideas in this chapter

| §    | Topic                | Idea                                       |
|------|----------------------|--------------------------------------------|
| 9.1  | Regions of the plane | What changes when you add one more         |
| 9.2  | Ramsey theory        | The pigeonhole principle                   |
| 9.3  | Graph colouring      | Coupled recurrences; chromatic polynomials |
| 9.4  | Generating functions | A sequence becomes a function              |
| 9.5  | Partitions           | Counting via coefficients                  |
| 9.6  | Sicherman dice       | Factorising polynomials                    |
| 9.7  | Thue–Morse           | Repeated roots give balanced sums          |
| 9.8  | Derangements         | Inclusion–exclusion; $1/e$                 |
| 9.9  | Coincidences         | Complements; the square-root law           |
| 9.10 | The ballot problem   | The reflection principle                   |
| 9.11 | Coupon collector     | Linearity of expectation; $n \ln n$        |
| 9.12 | Random subsets       | One element at a time                      |
| 9.13 | Loaded dice          | Impossibility by AM–GM                     |

**How to read this chapter**

Always start with small cases. Draw three lines, list the six orders of three cards, count the ways to make 5p. Small cases give you data to guess from, and they check your formulas at the end. But do not trust a pattern until you have a reason for it: Section 9.1 includes a famous sequence,  $1, 2, 4, 8, 16, \dots$ , whose next term is 31.



# 9.1 Lines Cutting the Plane

## The question

A plane contains  $n$  distinct given lines, no two of which are parallel, and no three of which intersect at a point. By first considering the cases  $n = 1, 2, 3$  and 4, provide and justify, by induction or otherwise, a formula for the number of line segments (including the infinite segments).

Prove also that the plane is divided into  $\frac{1}{2}(n^2 + n + 2)$  regions (including those extending to infinity).

## The hook

How many pieces can you cut a pizza into with  $n$  straight cuts, if the pieces need not be the same size? One cut gives 2, two cuts give 4, three cuts give 7 (not 8: try it), four give 11. The sequence 1, 2, 4, 7, 11, 16, ... is known as the *lazy caterer's sequence*, and this question asks for its formula.

The question is a first lesson in a technique that runs through all of combinatorics: do not try to count everything at once. Instead, ask what changes when you add one more object. And the answer turns out to be a special case of Euler's formula  $V - E + F = 2$ , which you met with the Platonic solids in Section 7.9.

## Before you start



**Pause and try** Draw the cases  $n = 1, 2, 3, 4$  carefully, keeping the lines in general position (no two parallel, no three through a point). Count the segments and the regions in each case. Can you guess the patterns?

You should find:

| $n$      | 0 | 1 | 2 | 3 | 4  |
|----------|---|---|---|---|----|
| segments | 0 | 1 | 4 | 9 | 16 |
| regions  | 1 | 2 | 4 | 7 | 11 |

The segments look like  $n^2$ . The regions go up by 1, 2, 3, 4: the differences are the natural numbers.

## Working it through

### STAGE 1 Counting segments



**Ask yourself** Concentrate on one line. How many of the other lines cross it, and into how many pieces do they cut it?

Each line is crossed by each of the other  $n - 1$  lines (none is parallel to it), and at  $n - 1$  different points (no three lines meet at a point). Those  $n - 1$  points cut the line into  $n$  segments: two infinite ones at the ends, and  $n - 2$  finite ones between. With  $n$  lines, there are

$$n \times n = n^2 \text{ segments.} \quad \blacksquare$$

This is a direct count, with no induction needed: we counted segments line by line.

### STAGE 2 What happens when we add a line?



**Ask yourself** Suppose  $n - 1$  lines are drawn, and you add the  $n$ th. It is cut into  $n$  segments by the others. What does each segment do to the region it passes through?

As the new line is drawn, each of its  $n$  segments runs through one of the existing regions and splits it into two (Figure 9.1.1). Different segments lie in different regions, because to pass from one region to another the line must cross an existing line. So adding the  $n$ th line increases the number of regions by exactly  $n$ :

$$R(n) = R(n - 1) + n.$$

### STAGE 3 Solving the recurrence

Starting from  $R(0) = 1$  (no lines, one region),

$$R(n) = 1 + (1 + 2 + \cdots + n) = 1 + \frac{n(n+1)}{2} = \frac{n^2 + n + 2}{2}.$$

To present this as an induction: the formula holds for  $n = 1$ , since  $\frac{1}{2}(1 + 1 + 2) = 2$ ; and if  $R(n - 1) = \frac{1}{2}((n - 1)^2 + (n - 1) + 2) = \frac{1}{2}(n^2 - n + 2)$ , then  $R(n) = \frac{1}{2}(n^2 - n + 2) + n = \frac{1}{2}(n^2 + n + 2)$ .  $\blacksquare$

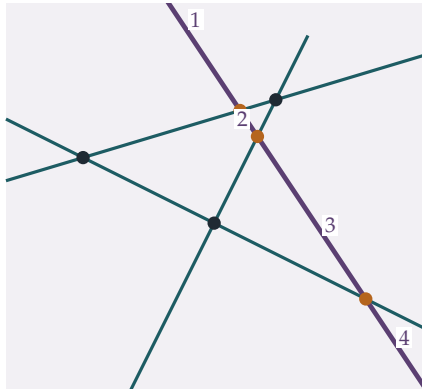


**Figure 9.1.1** Three lines (teal) divide the plane into 7 regions. The fourth line (plum) is crossed by the other three at the amber points, which cut it into four segments, numbered 1 to 4. Each segment splits one region into two, so the fourth line adds 4 regions, giving 11.

## Looking back

- **Count one piece at a time.** Segments were counted line by line; regions were counted by what one new line adds.
- **Turn a geometric step into a recurrence.** The observation “each new segment splits one region” became  $R(n) = R(n-1) + n$ .
- **General position matters.** Parallel lines, or three lines through a point, would give fewer regions. Find where each assumption was used.

## What’s really going on

### Euler’s formula in disguise

Here is a second proof, which explains the answer rather than just deriving it. Think of the arrangement of lines as a map: the crossing points are *vertices*, the segments are *edges* and the regions are *faces*. There are  $V = \binom{n}{2}$  vertices (one for each pair of lines) and  $E = n^2$  edges. The infinite segments have only one end, so add a single extra vertex “at infinity” where they all meet; now the picture is a map drawn on a sphere (imagine the plane wrapped round a ball, with infinity at the top). Euler’s formula  $V - E + F = 2$  from Section 7.9 gives

$$\left(\binom{n}{2} + 1\right) - n^2 + F = 2, \quad \text{so} \quad F = n^2 - \frac{n(n-1)}{2} + 1 = \frac{n^2 + n + 2}{2}.$$

The same number, from a completely different direction. The regions also split neatly into types:  $2n$  of them are unbounded (one between each pair of neighbouring infinite rays), and the remaining  $\frac{1}{2}(n^2 + n + 2) - 2n = \binom{n-1}{2}$  are bounded.

### Binomial coefficients and higher dimensions

The formula can be written as

$$R(n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2},$$

and this is not a coincidence. Turn the picture so that no line is horizontal. A region that is bounded below has a lowest point, which must be a crossing, and each crossing is the lowest point of exactly one region (the one directly above it): that gives  $\binom{n}{2}$  regions, one for each pair of lines. Every other region stretches down for ever, and so meets a horizontal line drawn far below all the crossings; the  $n$  lines cut that horizontal line into  $n + 1$  intervals, one in each such region. In total,  $\binom{n}{2} + n + 1$ . In three dimensions,  $n$  planes in general position cut space into

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} = \frac{n^3 + 5n + 6}{6}$$

pieces (the *cake numbers* 1, 2, 4, 8, 15, 26, ...), and the pattern continues in every dimension. The proof is the same induction: a new plane is cut by the others into regions arranged like lines in a plane, and each of those regions splits a piece of space in two.

#### Steiner and the arrangement of hyperplanes

The formula for regions cut by lines and planes was found by Jakob Steiner in 1826. Today it is the starting point of the theory of *hyperplane arrangements*, which counts the regions formed by arbitrary collections of lines, planes and their higher-dimensional analogues, including non-general ones. A remarkable 1975 theorem of Thomas Zaslavsky counts the regions of any arrangement using a polynomial that encodes which hyperplanes meet. The same ideas appear in machine learning: a neural network built from “ReLU” units cuts its input space into regions by hyperplanes, and the number of regions measures how complicated a function it can represent.

### The thinking skill

#### Ask what changes when you add one more

*To count a configuration of  $n$  objects, compare it with the configuration of  $n - 1$ . The change is often much easier to count than the whole.*

This is induction used as a tool for discovery, not just for proof. It is one of the most common moves in combinatorics, and it turns a geometric question into a recurrence relation that can be solved.

### Try this next

---

- 1 **Circles.** Show that  $n$  circles, each pair meeting in two points and no three through a point, divide the plane into  $n^2 - n + 2$  regions.
- 2 **Parallel lines.** If the  $n$  lines include  $k$  that are all parallel to each other (and otherwise general position), how many regions are there?
- 3 **Cake numbers.** Prove by induction that  $n$  planes in general position divide space into  $\frac{1}{6}(n^3 + 5n + 6)$  regions.
- 4 **A pattern that breaks.** Put  $n$  points on a circle and join every pair with a chord, with no three chords through one point inside the circle. For  $n = 1, 2, 3, 4, 5$  you get 1, 2, 4, 8, 16 regions. Count the regions for  $n = 6$ . (*Hint:* use Euler's formula; each set of four points gives one crossing.) Show that the number is  $1 + \binom{n}{2} + \binom{n}{4}$ .

### Further reading

---

- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics*** (Addison-Wesley), Chapter 1. Starts with exactly this problem, and the “pizza” and “zig” variants.
- **Miklós Bóna, *A Walk Through Combinatorics*** (World Scientific). A friendly first course, with chapters on induction, graphs and Euler's formula.
- **Richard Guy, “The strong law of small numbers”, *American Mathematical Monthly* 95 (1988).** Many patterns that look obvious and turn out to be false, including the circle problem in the exercises.



## 9.2 Ramsey's Theorem in Miniature

### The question

Six points  $A, B, C, D, E$  and  $F$  lie in three dimensional space and are in general positions, that is, no three are collinear and no four lie on a plane. All possible line segments joining pairs of points are drawn and coloured either gold or silver. Prove that there is a triangle whose edges are entirely of one colour. [Hint: consider segments radiating from  $A$ .]

Give a sketch showing that the result is false for five points in general positions.

### The hook

At any party of six people, there are either three who all know each other or three who are all strangers to each other. This is guaranteed, whoever the six people are. With five people it is not guaranteed. This little fact is the first case of *Ramsey's theorem*, and it captures a principle that runs through modern combinatorics: *complete disorder is impossible*. Any sufficiently large structure, however it is coloured or arranged, must contain a highly ordered piece.

The proof takes four lines. But the obvious next question (how many people guarantee four mutual friends or four mutual strangers? Five?) leads to some of the hardest open problems in mathematics.

### Before you start



**Ask yourself** The six points matter only through which pairs are joined by which colour: the geometry is a distraction. How many segments are there, and how many leave each point?

There are  $\binom{6}{2} = 15$  segments, and 5 leave each point. (The general-position condition just makes sure every three points form a genuine triangle. The problem is really about the *complete graph* on six vertices,  $K_6$ , with its edges coloured in two colours.)

### Working it through

#### STAGE 1 Five segments, two colours

**Ask yourself** Five segments leave  $A$ , each gold or silver. What can you say for certain about their colours?

At least three of them have the same colour. If at most two were gold and at most two silver, there would be at most four segments. This is the *pigeonhole principle*: five objects in two boxes put at least three in one box. Say (swapping the names of the colours if necessary) that  $AB$ ,  $AC$  and  $AD$  are all gold.

## STAGE 2 Look at the far ends

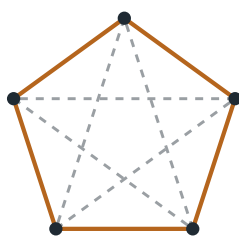
**Ask yourself** Now look at the triangle  $BCD$ . What happens if any one of its sides is gold? What if none of them is?

If any side of triangle  $BCD$  is gold, say  $BC$ , then  $ABC$  has all three sides gold:  $AB$ ,  $AC$  and  $BC$ . If no side of  $BCD$  is gold, then  $BC$ ,  $CD$  and  $DB$  are all silver, and  $BCD$  is a silver triangle. Either way there is a one-colour triangle (Figure 9.2.1, right). ■

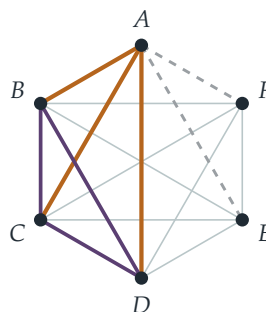
## STAGE 3 Five points are not enough

**Ask yourself** With five points, each point has only four segments. How might you colour them so that each point has exactly two of each colour?

Place the five points as a regular pentagon (lifting them slightly off a plane if you want general position). Colour the five sides of the pentagon gold and the five diagonals silver (Figure 9.2.1, left). The gold edges form a five-cycle, which contains no triangle; the silver edges form a pentagram, which is also a five-cycle and contains no triangle. So there is no one-colour triangle. ■



five points: no one-colour triangle



six points: look at  $BCD$

**Figure 9.2.1** Left: five points with the pentagon gold (solid) and the pentagram silver (dashed) have no one-colour triangle. Right: with six points, three of the five segments from  $A$  share a colour, say gold to  $B$ ,  $C$ ,  $D$ . The triangle  $BCD$  (plum) then decides: one gold side makes a gold triangle with  $A$ , and no gold side makes  $BCD$  a silver triangle.

## Looking back

---

- ▶ **Strip away the irrelevant.** The points in space became a coloured graph; only the colours of the edges mattered.
- ▶ **Pigeonhole.** Five edges and two colours forced three of one colour.
- ▶ **Case analysis on the right object.** Once three gold edges were found, the triangle at their far ends settled everything.
- ▶ **Sharpness.** The pentagon shows the number six cannot be lowered. A good theorem comes with an example showing it is best possible.

## What's really going on

---

### Ramsey numbers

The *Ramsey number*  $R(s, t)$  is the smallest  $N$  such that every colouring of the edges of  $K_N$  in gold and silver contains either a gold  $K_s$  (a set of  $s$  points with all edges between them gold) or a silver  $K_t$ . The question has shown that  $R(3, 3) = 6$ : six points always work, and the pentagon shows five do not.

Frank Ramsey proved in 1928 that these numbers always exist: for every  $s$  and  $t$  there is some  $N$  that works. The proof is an induction that generalises the argument above, and shows  $R(s, t) \leq R(s - 1, t) + R(s, t - 1)$ . But exact values are extraordinarily hard to find:

$$R(3, 3) = 6, \quad R(4, 4) = 18, \quad 43 \leq R(5, 5) \leq 46.$$

The value of  $R(4, 4)$  was found in 1955. Despite enormous computer searches,  $R(5, 5)$  is still unknown; the upper bound of 46 was proved only in 2024. A computer cannot simply check every colouring of  $K_{43}$ : there are  $2^{903}$  of them, far more than the number of atoms in the universe.

### Erdős and the aliens

Paul Erdős, who did more than anyone to develop Ramsey theory, liked to tell this story. Suppose aliens invade the Earth and threaten to destroy it unless we tell them the value of  $R(5, 5)$ . We should marshal all our computers and mathematicians, and we might find it. But if they asked for  $R(6, 6)$ , we should try to destroy the aliens. Erdős also proved, in 1947, that  $R(k, k) > 2^{k/2}$ , using a startling method: he showed that a *randomly* coloured graph on fewer than  $2^{k/2}$  points has a positive probability of containing no one-colour  $K_k$ , so a good colouring must exist, even though nobody knows how to construct one. This *probabilistic method* is now one of the most powerful tools in combinatorics.

---

### Order from chaos

Ramsey's theorem has many relatives, all expressing the same idea: large enough structures cannot avoid order. *Van der Waerden's theorem* says that if the whole numbers are

coloured with finitely many colours, there are arbitrarily long arithmetic progressions of one colour. The *Erdős–Szekeres theorem* says that any sequence of  $n^2 + 1$  distinct numbers contains an increasing or a decreasing subsequence of length  $n + 1$ . And Ramsey himself was interested in a question of logic: he proved his theorem as a lemma in a paper on deciding the truth of logical statements.

## The thinking skill

### The pigeonhole principle

*If more than  $kn$  objects are placed in  $n$  boxes, some box contains more than  $k$  objects.*

The principle is obvious, but using it well is an art: the skill is in choosing what the objects and boxes should be. Here they were the five edges at one point and the two colours. At university, pigeonhole arguments appear in number theory (approximating irrationals by fractions), in analysis (the Bolzano–Weierstrass theorem) and throughout combinatorics.

## Try this next

- 1 **At least two.** Show that in any two-colouring of  $K_6$  there are in fact at least *two* one-colour triangles. (*Hint:* count the “bichromatic angles”: pairs of edges of different colours at a common vertex. A point with  $g$  gold and  $5 - g$  silver edges is at the apex of  $g(5 - g) \leq 6$  such angles. Each triangle that is not one-colour contains exactly two.)
- 2 **Three colours.** If the edges of  $K_{17}$  are coloured in three colours, show that there is a one-colour triangle. (*Hint:* one point has at least six edges of one colour.)
- 3 **Three or four.** Show that  $R(3, 4) \leq 10$ : in any gold/silver colouring of  $K_{10}$  there is a gold triangle or a silver  $K_4$ . (*Hint:* a point has either at least 4 gold edges or at least 6 silver edges.) In fact  $R(3, 4) = 9$ .
- 4 **Monotone subsequences.** Prove the Erdős–Szekeres theorem for  $n = 2$ : any sequence of 5 distinct numbers contains an increasing or decreasing subsequence of length 3. Find a sequence of 4 with neither.

## Further reading

- Miklós Bóna, *A Walk Through Combinatorics* (World Scientific), the chapters on the pigeonhole principle and Ramsey theory.
- Ronald Graham & Joel Spencer, “Ramsey theory”, *Scientific American* (July 1990). A popular introduction by two leading researchers.

- **Noga Alon & Joel Spencer, *The Probabilistic Method*** (Wiley). The standard book on Erdős's method; the first chapter is accessible.
- **Paul Hoffman, *The Man Who Loved Only Numbers*** (Fourth Estate). A biography of Erdős.



## 9.3 Colouring Fence Posts

### The question

I have  $n$  fence posts placed in a line and, as part of my spouse's birthday celebrations, I wish to paint them using three different colours red, white and blue in such a way that no adjacent fence posts have the same colours. (This allows the possibility of using fewer than three colours as well as exactly three.) Let  $r_n$  be the number of ways (possibly zero) that I can paint them if I paint the first and the last post red and let  $s_n$  be the number of ways that I can paint them if I paint the first post red but the last post either of the other two colours. Explain why  $r_{n+1} = s_n$  and find  $r_n + s_n$ . Hence find the value of  $r_{n+1} + r_n$  for all  $n \geq 1$ .

Prove, by induction, that

$$r_n = \frac{2^{n-1} + 2(-1)^{n-1}}{3}.$$

Find the number of ways of painting  $n$  fence posts (where  $n \geq 3$ ) placed in a circle using three different colours in such a way that no adjacent fence posts have the same colours.

### The hook

Painting posts in a line is easy: 3 choices for the first, then 2 for each of the rest (anything but the colour of the previous post), giving  $3 \times 2^{n-1}$ . Bend the line into a circle and it becomes surprisingly awkward, because the last post now has to differ from the first as well as from its neighbour, and whether that rules out one colour or none depends on everything that came before.

The question gets round this with a classic trick: split the count into cases according to the colour of the last post, and find how the cases feed into each other. The answer,  $2^n + 2(-1)^n$ , is the first example of a *chromatic polynomial*, part of the mathematics behind the four colour theorem.

## Before you start



**Pause and try** Count the ways to paint 3 posts in a circle, and 4. Does  $3 \times 2^{n-1}$  overcount, and by how much?

Three posts in a circle are all adjacent to each other, so they need three different colours:  $3! = 6$  ways. Four posts in a circle: 18 ways (you can check this at the end). Compare  $3 \times 2^2 = 12$  and  $3 \times 2^3 = 24$  for the line: the circle loses the colourings in which the last post matches the first.

## Working it through

### STAGE 1 Two kinds of colouring



**Ask yourself** Take a colouring of  $n + 1$  posts with the first and last red. Remove the last post. What can you say about the colouring of the first  $n$  posts that remains?

Post  $n$  is next to post  $n + 1$ , which is red, so post  $n$  is not red. Removing post  $n + 1$  leaves a colouring of  $n$  posts with the first red and the last not red: one of the  $s_n$  colourings. Conversely, any of the  $s_n$  colourings can be extended in exactly one way, by painting post  $n + 1$  red. So the two sets correspond exactly, and  $r_{n+1} = s_n$ .



**Ask yourself** What is  $r_n + s_n$ ?

Every colouring with the first post red has its last post either red or not, so  $r_n + s_n$  counts all colourings of  $n$  posts with the first red. After the first post, each post has 2 choices, so  $r_n + s_n = 2^{n-1}$ . Hence

$$r_{n+1} + r_n = s_n + r_n = 2^{n-1}. \quad \blacksquare$$

### STAGE 2 The formula by induction

For  $n = 1$  there is one post, painted red, so  $r_1 = 1$ , and the formula gives  $\frac{1}{3}(1 + 2) = 1$ . Suppose the formula holds for  $r_n$ . Then

$$r_{n+1} = 2^{n-1} - r_n = 2^{n-1} - \frac{2^{n-1} + 2(-1)^{n-1}}{3} = \frac{2 \cdot 2^{n-1} - 2(-1)^{n-1}}{3} = \frac{2^n + 2(-1)^n}{3},$$

which is the formula for  $r_{n+1}$ .  $\blacksquare$



**Pause and try** Where did the formula come from? The recurrence  $r_{n+1} + r_n = 2^{n-1}$  is like a differential equation: a “particular solution”  $c \cdot 2^n$  plus a

“complementary function”  $A(-1)^n$  that solves  $r_{n+1} + r_n = 0$ . Find  $c$  and  $A$ .

### STAGE 3 The circle



**Ask yourself** In a circle of  $n$  posts, fix the colour of post 1, say red. Now cut the circle between post  $n$  and post 1, and straighten it into a line. Which of the line colourings correspond to valid circle colourings?

Straightened out, the circle becomes a line of  $n$  posts starting with red, in which the last post must also differ from post 1: the last post is not red. These are exactly the  $s_n$  colourings. Post 1 can be any of 3 colours, so the number of circle colourings is

$$3s_n = 3(2^{n-1} - r_n) = 3 \cdot 2^{n-1} - 2^{n-1} - 2(-1)^{n-1} = 2^n + 2(-1)^n. \quad \blacksquare$$

Check:  $n = 3$  gives  $8 - 2 = 6$ , and  $n = 4$  gives  $16 + 2 = 18$ .

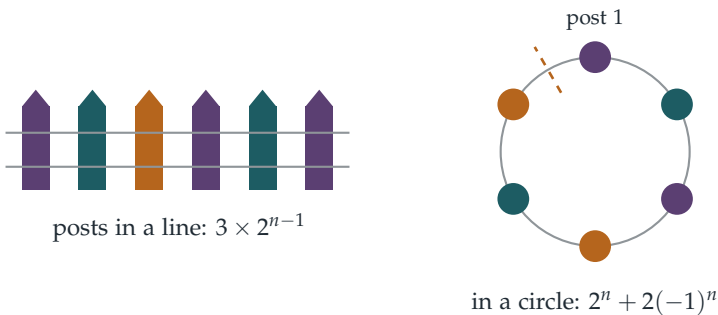


**Figure 9.3.1** Left: in a line, each post after the first just has to differ from its predecessor. Right: in a circle, the last post must also differ from post 1. Cutting the circle at the dashed line turns a circle colouring into a line colouring whose first and last posts differ: one of the  $s_n$  colourings counted in the question.

### Looking back

- **Split into cases that feed each other.** Recording whether the last post was red gave two sequences,  $r_n$  and  $s_n$ , linked by simple recurrences.
- **Bijections.**  $r_{n+1} = s_n$  was proved by matching colourings one-to-one, not by calculation.
- **Cut the circle.** A cyclic problem became a linear one with a condition on the ends.

### What's really going on

### Chromatic polynomials

The posts and their adjacencies form a *graph*: a path for the line, a cycle for the circle. A *proper colouring* gives adjacent vertices different colours. The number of proper colourings with  $k$  colours is a polynomial in  $k$ , the *chromatic polynomial* of the graph, introduced by George Birkhoff in 1912. Repeating the question's argument with  $k$  colours gives

$$\text{path: } k(k-1)^{n-1}, \quad \text{cycle: } (k-1)^n + (-1)^n(k-1).$$

With  $k = 3$  the cycle formula is  $2^n + 2(-1)^n$ . Birkhoff hoped to prove the four colour theorem (every map can be coloured with four colours so that neighbouring countries differ) by showing that the chromatic polynomial of every map's graph is positive at  $k = 4$ . That approach never succeeded; the theorem was finally proved in 1976 by Kenneth Appel and Wolfgang Haken, with extensive computer help. Chromatic polynomials remain an active field, with surprising connections to statistical physics, where they appear as the *Potts model* of magnetism.

### The same count by matrices

There is another way to see the answer, using linear algebra from Chapter 7. Let  $\mathbf{M}$  be the  $3 \times 3$  matrix with 0 on the diagonal and 1 elsewhere: entry  $(i, j)$  is 1 when colour  $j$  may follow colour  $i$ . Then entry  $(i, j)$  of  $\mathbf{M}^{n-1}$  counts line colourings that start with colour  $i$  and end with colour  $j$ , so  $r_n$  is a diagonal entry of  $\mathbf{M}^{n-1}$ . A circle colouring is a sequence that returns to its start after  $n$  steps, so the number of circle colourings is the *trace* of  $\mathbf{M}^n$ . Since  $\mathbf{M} = \mathbf{J} - \mathbf{I}$ , where every entry of  $\mathbf{J}$  is 1, the eigenvalues of  $\mathbf{M}$  are  $3 - 1 = 2$  (eigenvector  $(1, 1, 1)$ ) and  $-1$  (twice), and the trace of  $\mathbf{M}^n$  is the sum of the  $n$ th powers of the eigenvalues:

$$\text{tr } \mathbf{M}^n = 2^n + (-1)^n + (-1)^n = 2^n + 2(-1)^n.$$

This *transfer matrix* method is used throughout combinatorics and statistical physics to count arrangements built up one step at a time.

### The thinking skill

#### Set up coupled recurrences

*When one sequence will not satisfy a simple recurrence on its own, record a little extra information (here, the colour of the last post) and find recurrences linking the cases.*

This is the idea of a *state*: keep track of exactly what you need to know to continue, and nothing more. It is the basis of dynamic programming in computer science, of Markov chains in probability (Section 12.16), and of the transfer matrix method.

### Try this next

---

- 1  **$k$  colours.** Repeat the question with  $k$  colours. Show that  $r_{n+1} = s_n$  still holds and that  $r_{n+1} + r_n = (k-1)^{n-1}$ , find  $r_n$ , and deduce the chromatic polynomial of the cycle,  $(k-1)^n + (-1)^n(k-1)$ .
- 2 **No two reds together.** How many ways are there to paint  $n$  posts in a line red or white, with no two adjacent posts red? (You should find the Fibonacci numbers.)
- 3 **A wheel.** Add a central post adjacent to every post of the circle (a “wheel”). How many ways are there to colour it with 4 colours? (*Hint:* the centre uses one colour, and the circle must use the other three.)
- 4 **Necklaces.** If two colourings of the circle count as the same when one is a rotation of the other, how many different colourings are there for  $n = 4$ ? For  $n = 5$ ? (This is the start of *Burnside’s lemma*.)

### Further reading

---

- **Robin Wilson, *Introduction to Graph Theory*** (Pearson). The standard first course; covers colourings and the four colour problem.
- **Robin Wilson, *Four Colours Suffice*** (Penguin). The story of the four colour theorem, for a general reader.
- **Herbert Wilf, *generatingfunctionology*** (free online). Recurrences and how to solve them, including the transfer matrix method.



## 9.4 Sequences as Functions

### The question

Given an infinite sequence of numbers  $u_0, u_1, u_2, \dots$ , we define the *generating function*,  $f$ , for the sequence by

$$f(x) = u_0 + u_1x + u_2x^2 + u_3x^3 + \dots$$

Issues of convergence can be ignored in this question.

- (i) Using the binomial series, show that the sequence given by  $u_n = n$  has generating function  $x(1-x)^{-2}$ , and find the sequence that has generating function  $x(1-x)^{-3}$ .

Hence, or otherwise, find the generating function for the sequence  $u_n = n^2$ . You should simplify your answer.

- (ii) (a) The sequence  $u_0, u_1, u_2, \dots$  is determined by  $u_n = ku_{n-1}$  ( $n \geq 1$ ), where  $k$  is independent of  $n$ , and  $u_0 = a$ . By summing the identity  $u_n x^n \equiv ku_{n-1}x^n$ , or otherwise, show that the generating function,  $f$ , satisfies

$$f(x) = a + kx f(x).$$

Write down an expression for  $f(x)$ .

- (b) The sequence  $u_0, u_1, u_2, \dots$  is determined by  $u_n = u_{n-1} + u_{n-2}$  ( $n \geq 2$ ) and  $u_0 = 0, u_1 = 1$ . Obtain the generating function.

### The hook

Work out  $1/89$  on a calculator: 0.011235955... The digits begin 0, 1, 1, 2, 3, 5, the Fibonacci numbers. Is this a coincidence? It is not, and this question explains why.

The idea is one of the most powerful in combinatorics. Take an infinite sequence of numbers and package it as a single function, by using the terms as the coefficients of a power series. Then facts about the sequence become facts about the function, and you can use algebra and calculus on them. Herbert Wilf put it memorably: "A generating function is a clothesline on which we hang up a sequence of numbers for display."

## Before you start

**Ask yourself** What sequence has generating function  $\frac{1}{1-x}$ ? What about  $\frac{1}{1-2x}$ ? What happens to the sequence if you multiply its generating function by  $x$ ?

The geometric series  $\frac{1}{1-x} = 1 + x + x^2 + \dots$  is the generating function of  $1, 1, 1, \dots$ , and  $\frac{1}{1-2x} = 1 + 2x + 4x^2 + \dots$  that of the powers of 2. Multiplying by  $x$  shifts every term one place along:  $x \cdot \frac{1}{1-x} = x + x^2 + \dots$  belongs to  $0, 1, 1, 1, \dots$

## Working it through

### STAGE 1 The natural numbers and the triangle numbers

**Ask yourself** Expand  $(1-x)^{-2}$  by the binomial series. What is the coefficient of  $x^n$ ?

The binomial series gives

$$(1-x)^{-2} = 1 + 2x + 3x^2 + 4x^3 + \dots = \sum_{n \geq 0} (n+1)x^n,$$

so  $x(1-x)^{-2} = x + 2x^2 + 3x^3 + \dots = \sum nx^n$ : the generating function of  $u_n = n$ .

Similarly, the coefficient of  $x^n$  in  $(1-x)^{-3}$  is  $\frac{(-3)(-4)\dots(-3-n+1)}{n!}(-1)^n = \frac{(n+1)(n+2)}{2}$ , so

$$x(1-x)^{-3} = \sum_{n \geq 1} \frac{n(n+1)}{2} x^n = x + 3x^2 + 6x^3 + 10x^4 + \dots,$$

the generating function of the triangle numbers  $u_n = \frac{1}{2}n(n+1)$ .

**Ask yourself** How can you write  $n^2$  in terms of  $n$  and  $\frac{1}{2}n(n+1)$ ?

Since  $n^2 = 2 \cdot \frac{1}{2}n(n+1) - n$ , and generating functions add just as sequences do,

$$\sum n^2 x^n = \frac{2x}{(1-x)^3} - \frac{x}{(1-x)^2} = \frac{2x - x(1-x)}{(1-x)^3} = \frac{x(1+x)}{(1-x)^3}. \quad \blacksquare$$

### STAGE 2 A recurrence becomes an equation

**Ask yourself** Multiply  $u_n = ku_{n-1}$  by  $x^n$  and add up over all  $n \geq 1$ . What is each side, in terms of  $f$ ?

The left side is  $\sum_{n \geq 1} u_n x^n = f(x) - u_0 = f(x) - a$ . The right side is  $kx \sum_{n \geq 1} u_{n-1} x^{n-1} = kx f(x)$ . So  $f(x) = a + kx f(x)$ , and

$$f(x) = \frac{a}{1 - kx}. \quad \blacksquare$$

Of course,  $u_n = ak^n$ ; expanding  $\frac{a}{1 - kx}$  as a geometric series confirms it.

### STAGE 3 The Fibonacci numbers

Do the same with  $u_n = u_{n-1} + u_{n-2}$ , for  $n \geq 2$ :

$$\sum_{n \geq 2} u_n x^n = x \sum_{n \geq 2} u_{n-1} x^{n-1} + x^2 \sum_{n \geq 2} u_{n-2} x^{n-2},$$

that is,

$$f(x) - u_0 - u_1 x = x(f(x) - u_0) + x^2 f(x).$$

With  $u_0 = 0$  and  $u_1 = 1$ , this says  $f(x) - x = x f(x) + x^2 f(x)$ , so

$$f(x) = \frac{x}{1 - x - x^2}. \quad \blacksquare$$

### STAGE 4 Explaining $1/89$



**Pause and try** Ignoring convergence (which is fine here: the series converges for  $|x| < 0.618$ ), put  $x = \frac{1}{10}$  in  $f$ . What do you get?

$$\frac{1}{10}F_1 + \frac{1}{100}F_2 + \frac{1}{1000}F_3 + \cdots = f\left(\frac{1}{10}\right) = \frac{1/10}{1 - 1/10 - 1/100} = \frac{10}{89}.$$

Dividing by 10,  $\frac{1}{89} = \sum F_n / 10^{n+1}$ . The Fibonacci numbers are placed in successive decimal places, overlapping when they get bigger than 9 (Figure 9.4.1). The early digits of  $1/89$  show the Fibonacci numbers; later ones are blurred by carrying.

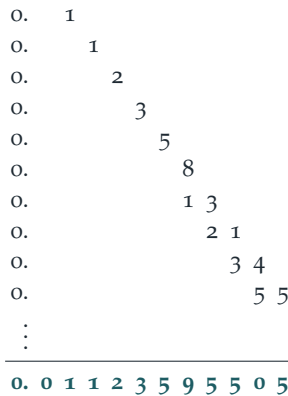


**Figure 9.4.1** Adding  $F_n / 10^{n+1}$  for  $n = 1, 2, 3, \dots$ . Each Fibonacci number ends one decimal place further to the right. Once they have two digits they overlap, and carrying changes the digits from the seventh place on:  
 $\frac{1}{89} = 0.01123595505 \dots$

## Looking back

---

- **Turn a sequence into a function.** The whole infinite sequence became one rational function.
- **Recurrences become equations.** Multiplying by  $x^n$  and summing turned the rule  $u_n = u_{n-1} + u_{n-2}$  into an equation for  $f$  that could be solved by algebra.
- **Build new from old.** Adding, shifting and differentiating generating functions produced new ones without starting from scratch.

## What's really going on

---

### From generating function to formula

Generating functions do not just encode a sequence: they can solve for it. The denominator factorises as  $1 - x - x^2 = (1 - \varphi x)(1 - \psi x)$ , where  $\varphi = \frac{1+\sqrt{5}}{2}$  and  $\psi = \frac{1-\sqrt{5}}{2}$  are the roots of  $t^2 = t + 1$ . Partial fractions give

$$\frac{x}{1 - x - x^2} = \frac{1}{\sqrt{5}} \left( \frac{1}{1 - \varphi x} - \frac{1}{1 - \psi x} \right),$$

and expanding each fraction as a geometric series,

$$F_n = \frac{\varphi^n - \psi^n}{\sqrt{5}}.$$

This is *Binet's formula*. Since  $|\psi| < 1$ , the second term is tiny, so  $F_n$  is the nearest integer to  $\varphi^n / \sqrt{5}$ , and the ratio of successive Fibonacci numbers tends to the golden ratio (the subject of Section 2.2). The same method solves any linear recurrence with constant coefficients; the exercises include an example from STEP I 2004.

### Why generating functions are so useful

The real power appears when we multiply. The coefficient of  $x^n$  in a product  $f(x)g(x)$  is  $\sum_k u_k v_{n-k}$ , which adds up over all ways of splitting  $n$  into two parts. That is exactly what we need to count things built from two pieces. The next two sections use it to count the ways of making change (Section 9.5) and to find strangely numbered dice with the same totals as ordinary ones (Section 9.6). In probability, the same idea gives the *probability generating function*, which turns sums of independent random variables into products.

### De Moivre, Euler and Laplace

Abraham de Moivre used generating functions in the 1720s to solve linear recurrences, essentially by the method above. Leonhard Euler used them in 1748 to study partitions of whole numbers, and Pierre-Simon Laplace made them the foundation of his *Théorie analytique des probabilités* (1812). In the twentieth century they became central to

combinatorics, and analytic methods (studying a generating function as a function of a complex variable) now give precise estimates for counting problems far too complicated to solve exactly.

## The thinking skill

### Change the representation

*A problem about a sequence may become easy when the sequence is turned into a function, a matrix, or a picture.*

Generating functions translate recurrences into algebra and counting into multiplication. At university you will meet many such translations: Fourier series turn functions into sequences, Laplace transforms turn differential equations into algebra, and matrices turn geometry into arithmetic.

## Try this next

- 1 Two digits at a time.** (From STEP II, 2022.) Using  $F_r \leq 2^{r-n} F_n$  for  $r \geq n$ , show that  $\sum_{r \geq 7} F_r / 10^r < 2 \times 10^{-6}$ , and hence find the first six digits after the decimal point of  $1/89$  with justification. Then find a fraction  $r/s$ , with  $r$  and  $s$  positive integers less than 10000, whose decimal expansion starts 0.0001010203050813213455...
- 2 Solving a recurrence.** (From STEP I, 2004.) Let  $a_0 = 2$ ,  $a_1 = 7$  and  $a_n = 7a_{n-1} - 10a_{n-2}$ . Show that the generating function is  $\frac{1}{1-2x} + \frac{1}{1-5x}$ , and hence that  $a_n = 2^n + 5^n$ . If  $b_0 = 5$ ,  $b_1 = 10$ ,  $b_2 = 40$ ,  $b_3 = 100$  and  $b_n = pb_{n-1} + qb_{n-2}$ , find  $p$ ,  $q$  and  $b_n$ .
- 3 Counting numbers.** Show that  $\frac{1}{9801} = 0.0001020304\dots$ , with the two-digit blocks counting upwards. Where does the pattern first go wrong, and why?
- 4 Sums of squares.** Explain why  $\frac{1}{1-x}f(x)$  is the generating function of the partial sums  $u_0 + u_1 + \dots + u_n$ . Use this and the answer to part (i) to find a formula for  $1^2 + 2^2 + \dots + n^2$ .

## Further reading

- **Herbert Wilf, *generatingfunctionology*** (A K Peters; free online). The classic introduction. Chapter 1 is perfect after this section.
- **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics*** (Addison-Wesley), Chapter 7.
- **Philippe Flajolet & Robert Sedgewick, *Analytic Combinatorics*** (Cambridge University

Press; free online). Far beyond first year, but the opening pages show where the subject leads.

# 9.5 Making Change

## The question

Show that you can make up 10 pence in eleven ways using 10p, 5p, 2p and 1p coins.

In how many ways can you make up 20 pence using 20p, 10p, 5p, 2p and 1p coins?

[You are reminded that no credit will be given for unexplained answers.]

## The hook

This looks like the easiest question in the book: just list the possibilities. But listing gets out of hand very quickly. There are 41 ways to make 20p, 451 ways to make 50p, and 4563 ways to make £1 from the coins in a British pocket. The question is really asking for a *system*, a way to organise the count so that nothing is missed and nothing counted twice.

Behind the system is a beautiful piece of mathematics. Each coin corresponds to a geometric series, the number of ways to make  $n$  pence is a coefficient in their product, and the same idea, applied to all whole numbers at once, gives Euler's theory of *partitions*, one of the deepest topics in number theory.

## Before you start



**Ask yourself** How many ways are there to make  $n$  pence using only 2p and 1p coins? Why does the answer depend only on how many 2p coins you use?

Once the number of 2p coins is chosen, the rest must be 1p coins, so there is exactly one way for each possible number of 2p coins:  $0, 1, \dots, \lfloor n/2 \rfloor$ . That gives  $\lfloor n/2 \rfloor + 1$  ways. Call this  $a(n)$ . For example,  $a(10) = 6$  and  $a(5) = 3$ .

## Working it through

### STAGE 1 Ten pence



**Ask yourself** Organise the count by the largest coins. How many 10p coins can you use? Given that, how many 5p coins?

- ▶ One 10p coin: 1 way.
- ▶ No 10p, two 5p: 1 way.
- ▶ No 10p, one 5p: the other 5p is made from 2p and 1p, in  $a(5) = 3$  ways.
- ▶ No 10p or 5p: 10p from 2p and 1p, in  $a(10) = 6$  ways.

The total is  $1 + 1 + 3 + 6 = 11$ . ■

**STAGE 2 Building up layer by layer**



**Ask yourself** Let  $b(n)$  be the number of ways to make  $n$  pence from 5p, 2p and 1p coins. How can you express  $b(n)$  using the function  $a$ ?

Choose the number  $k$  of 5p coins; the remaining  $n - 5k$  pence are made from 2p and 1p coins. So

$$b(n) = a(n) + a(n - 5) + a(n - 10) + \dots$$

This gives

| $n$                      | 0 | 5 | 10 | 15 | 20 |
|--------------------------|---|---|----|----|----|
| $a(n)$                   | 1 | 3 | 6  | 8  | 11 |
| $b(n) = a(n) + b(n - 5)$ | 1 | 4 | 10 | 18 | 29 |

Similarly, let  $c(n)$  count ways using 10p coins as well; then  $c(n) = b(n) + b(n - 10) + \dots$ , so  $c(10) = 10 + 1 = 11$  (agreeing with Stage 1) and  $c(20) = b(20) + b(10) + b(0) = 29 + 10 + 1 = 40$ .

**STAGE 3 Twenty pence**

Finally the 20p coin: either we use it (one way) or we do not ( $c(20)$  ways). So there are

$$40 + 1 = 41 \text{ ways.} \quad \blacksquare$$



**Pause and try** Notice the structure: each new coin adds a layer, and each layer is a sum of the previous layer at spacings of the new coin's value. Can you see how a computer would carry out this calculation for £1, or £100?

**Looking back**

- ▶ **Organise by the largest part.** Fixing the number of the largest coins first made the count systematic.

- **Build a table.** Each new coin added a row computed from the row before. This is *dynamic programming*, the standard way computers count such things.
- **Check against a known case.** The table reproduced the 11 ways for 10p.

## What's really going on

### Coins as geometric series

Here is the idea that turns this into algebra. The number of 5p coins used can be  $0, 1, 2, \dots$ , contributing  $0, 5, 10, \dots$  pence. Record this as the series

$$1 + x^5 + x^{10} + x^{15} + \dots = \frac{1}{1 - x^5}.$$

Do the same for each coin and multiply:

$$\frac{1}{(1 - x)(1 - x^2)(1 - x^5)(1 - x^{10})(1 - x^{20})} = \sum_n w(n)x^n.$$

When the brackets are multiplied out, each term  $x^n$  arises from choosing one term from each bracket, that is, a number of each coin, with total value  $n$ . So  $w(n)$ , the coefficient of  $x^n$ , is exactly the number of ways to make  $n$  pence. The table in Stage 2 was simply multiplying out this product one bracket at a time. This is the generating function idea of Section 9.4, and multiplication is where it shows its power: products of generating functions count combinations of choices.

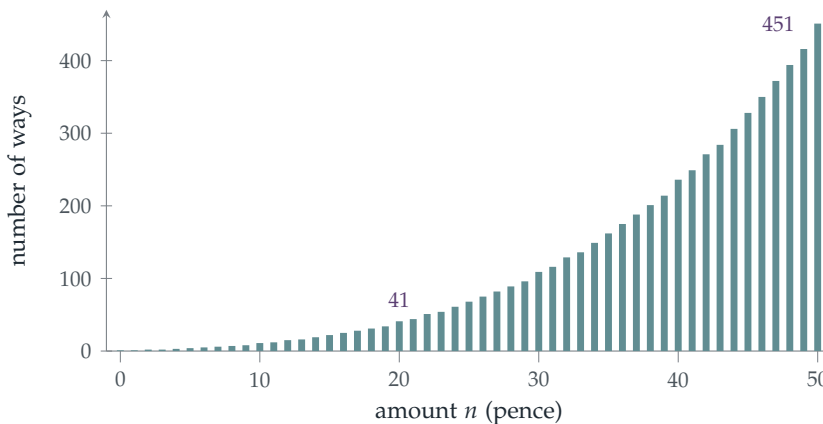


**Figure 9.5.1** The number of ways to make  $n$  pence from 1p, 2p, 5p, 10p, 20p and 50p coins. The jumps at multiples of 5 and 10 show where a new coin first becomes usable. With six coins the count grows roughly like  $n^5$ : in general, like a polynomial of degree one less than the number of coins.

### Partitions

What if every whole number is allowed as a coin? Then the number of ways to write  $n$  as a sum of positive integers, ignoring order, is the *partition number*  $p(n)$ . For example

$p(4) = 5$ , since  $4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1$ . Euler saw in 1748 that

$$\sum_{n \geq 0} p(n)x^n = \frac{1}{(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots},$$

and used it to discover remarkable identities. One of them: the number of ways to write  $n$  as a sum of *distinct* parts equals the number of ways to write it as a sum of *odd* parts. (The proof is a single line of algebra with generating functions: try it in the exercises.)

### Hardy, Ramanujan and $p(200)$

The partition numbers grow fast:  $p(10) = 42$ ,  $p(100) = 190\,569\,292$ . In 1918 G. H. Hardy and Srinivasa Ramanujan found an astonishing formula for  $p(n)$ , an infinite series whose first term is

$$p(n) \approx \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{2n/3}}.$$

To test it, Percy MacMahon computed  $p(200) = 3\,972\,999\,029\,388$  by hand, using Euler's identities. The first few terms of Hardy and Ramanujan's series gave the right answer to within a fraction of one. Ramanujan also discovered that  $p(5n+4)$  is always divisible by 5,  $p(7n+5)$  by 7, and  $p(11n+6)$  by 11. Explaining why there are no more such patterns took until the twenty-first century.

## The thinking skill

### Count via coefficients

*When an object is built from independent choices, each choice is a factor, and the number of objects of total size  $n$  is the coefficient of  $x^n$  in the product.*

This turns counting into algebra. At university it is the basis of enumerative combinatorics, and it is how probability generating functions handle sums of independent random variables.

## Try this next

- Two kinds of six.** (From STEP II, 2012.) Show that the coefficient of  $x^{24}$  in  $(1-x^6)^{-2}(1-x^3)^{-1}$  is 15, and find the coefficient of  $x^n$  in general. Then show that the coefficient of  $x^{24}$  in  $(1-x^6)^{-2}(1-x^3)^{-1}(1-x)^{-1}$  is 55, and find the coefficients of  $x^{25}$  and  $x^{66}$ . Describe a coin-counting problem that these coefficients answer.
- Distinct and odd.** Show that  $(1+x)(1+x^2)(1+x^3)\cdots$  is the generating function for partitions into distinct parts. Use  $1+x^k = \frac{1-x^{2k}}{1-x^k}$  to show that it equals  $\frac{1}{(1-x)(1-x^3)(1-x^5)\cdots}$ , and deduce Euler's theorem. Check it for  $n = 6$ .

- 3 A formula.** Show that the number of ways to make  $n$  pence using only 1p, 2p and 5p coins is the nearest integer to  $\frac{(n+4)^2}{20}$ . (Check it first for small  $n$ .)
- 4 Order matters.** In how many ways can you make  $n$  pence by putting 1p and 2p coins into a slot one at a time, if the order matters? (You have met this sequence before.)

### Further reading

---

- **George Andrews & Kimmo Eriksson, *Integer Partitions*** (Cambridge University Press). A short, readable introduction for undergraduates.
- **George Pólya, “On picture-writing”, *American Mathematical Monthly* 63 (1956).** Pólya’s classic explanation of generating functions using the problem of changing a dollar (there are 292 ways).
- **Robert Kanigel, *The Man Who Knew Infinity*** (Abacus). The life of Ramanujan, and his partnership with Hardy.



# 9.6 Strange Dice

## The question

I have two dice whose faces are all painted different colours. I number the faces of one of them 1, 2, 2, 3, 3, 6 and the other 1, 3, 3, 4, 5, 6. I can now throw a total of 3 in two different ways using the two number 2's on the first die once each. Show that there are seven different ways of throwing a total of 6.

I now renumber the dice (again only using integers in the range 1 to 6) with the results shown in the following table.

|                                       |   |   |   |   |   |   |   |   |    |    |    |
|---------------------------------------|---|---|---|---|---|---|---|---|----|----|----|
| Total shown by the two dice           | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Different ways of obtaining the total | 0 | 2 | 1 | 1 | 4 | 3 | 8 | 6 | 5  | 6  | 0  |

Find how I have numbered the dice explaining your reasoning.

[You will only get high marks if the examiner can follow your argument.]

## The hook

Is there any other way to number two six-sided dice, with positive whole numbers, so that every total from 2 to 12 comes up exactly as often as with ordinary dice? In 1978 Martin Gardner announced in *Scientific American* that there is exactly one: a die numbered 1, 2, 2, 3, 3, 4 and a die numbered 1, 3, 4, 5, 6, 8. They had been discovered by Colonel George Sicherman, and are now sold as *Sicherman dice*. A game of Monopoly played with them would be indistinguishable from an ordinary one.

This question is a warm-up: deduce how two dice are numbered from the table of totals. The systematic way to do it, and the proof that Sicherman's dice are the only strange pair, both come from turning each die into a polynomial.

## Before you start



**Ask yourself** If the first die has  $p_i$  faces showing  $i$  and the second has  $q_j$  faces showing  $j$ , how many ways are there to throw a total of  $t$ ?

The number of ways is  $\sum_i p_i q_{t-i}$ : choose the number  $i$  on the first die, then the second must show  $t - i$ . For the first pair of dice, a total of 6 comes from  $1 + 5$  (in  $1 \times 1$  way),  $2 + 4$  ( $2 \times 1$ ),  $3 + 3$  ( $2 \times 2$ ), and nothing else, since the first die has no 4 or 5 and the second has no 0. That is  $1 + 2 + 4 = 7$  ways. ■

## Working it through

### STAGE 1 The extremes

Call the dice  $A$  and  $B$ , and write  $a_i, b_i$  for the number of faces of each showing  $i$ .

**?** **Ask yourself** There is no way to throw 2, nor 12. What does that say about the 1s and the 6s? There *are* two ways to throw 3. Which die has the 1s?

No total of 2 means  $a_1b_1 = 0$ : at most one die has any 1s. A total of 3 can only be  $1 + 2$  or  $2 + 1$ , and there are two ways, so some die does have a 1; call it  $A$ , so  $b_1 = 0$ . Then

$$\text{total 3: } a_1b_2 = 2, \quad \text{so } (a_1, b_2) = (2, 1) \text{ or } (1, 2).$$

Similarly, no total of 12 means  $a_6b_6 = 0$ .

### STAGE 2 Working upwards

**?** **Ask yourself** Write down the equation for a total of 4, remembering  $b_1 = 0$ . What does it say in each of the two cases?

A total of 4 is  $1 + 3$  or  $2 + 2$  (and  $3 + 1$  is impossible as  $b_1 = 0$ ), so  $a_1b_3 + a_2b_2 = 1$ .

► **Case (a),**  $a_1 = 2, b_2 = 1$ :  $2b_3 + a_2 = 1$ , so  $b_3 = 0$  and  $a_2 = 1$ .

► **Case (b),**  $a_1 = 1, b_2 = 2$ :  $b_3 + 2a_2 = 1$ , so  $a_2 = 0$  and  $b_3 = 1$ .

Total 5:  $a_1b_4 + a_2b_3 + a_3b_2 = 1$ .

► **Case (a):**  $2b_4 + 0 + a_3 = 1$ , so  $b_4 = 0, a_3 = 1$ . So far  $A$  has 1, 1, 2, 3 and  $B$  has 2;  $B$ 's other five faces are 5s and 6s.

► **Case (b):**  $b_4 + 0 + 2a_3 = 1$ , so  $a_3 = 0, b_4 = 1$ . So far  $A$  has 1 and  $B$  has 2, 2, 3, 4;  $A$ 's other five faces are from 4, 5, 6, and  $B$ 's other two from 5, 6.

### STAGE 3 Case (a) fails

**?** **Ask yourself** In case (a), write down the equation for a total of 7, and use  $b_5 + b_6 = 5$ .

Total 7:  $a_1b_6 + a_2b_5 + a_3b_4 + a_4b_3 + a_5b_2 = 2b_6 + b_5 + 0 + 0 + a_5 = 3$ . With  $b_6 = 5 - b_5$  this says  $10 - b_5 + a_5 = 3$ , so  $b_5 = 7 + a_5 \geq 7$ . But  $B$  only has six faces. Case (a) is impossible.

### STAGE 4 Case (b) succeeds

**?** **Ask yourself** In case (b), use the totals 6 and 11, and the fact that  $a_6b_6 = 0$ .

Total 6:  $a_1b_5 + a_4b_2 = b_5 + 2a_4 = 4$  (the other products involve  $a_2 = a_3 = 0$  or  $b_1 = 0$ ). Since  $b_5 \leq 2$  and  $b_5 = 4 - 2a_4$  is even, either  $b_5 = 0, a_4 = 2$  or  $b_5 = 2, a_4 = 1$ .

- If  $b_5 = 0$  then  $B$  is 2, 2, 3, 4, 6, 6, so  $a_6 = 0$  and  $A$  is 1, 4, 4, 5, 5, 5. But then total 7 has  $a_1b_6 + a_4b_3 + a_5b_2 = 2 + 2 + 6 = 10$  ways, not 3.
- If  $b_5 = 2$  then  $B$  is 2, 2, 3, 4, 5, 5, with  $b_6 = 0$ , and  $A$  is 1, 4 and four faces from 5 and 6. Total 11:  $a_6b_5 = 2a_6 = 6$ , so  $a_6 = 3$  and  $a_5 = 1$ .

The only possibility is

$$A : 1, 4, 5, 6, 6, 6, \quad B : 2, 2, 3, 4, 5, 5.$$

Finally we must check that these dice give the whole table, since we have used only some of the totals. Multiplying out (or using the polynomials below) confirms every entry. ■

## Looking back

- **Start at the extremes.** The smallest and largest totals involve only one or two products, so they give the strongest information.
- **Split into cases and kill them.** Each total cut the possibilities down, until only one case survived.
- **Check sufficiency.** The deduction showed what the dice *must* be; a final check showed that they work.

## What's really going on

### Dice as polynomials

Write each die as a polynomial, with one term  $x^i$  for each face showing  $i$ . An ordinary die is  $x + x^2 + x^3 + x^4 + x^5 + x^6$ , and the dice in the question are

$$A(x) = x + x^4 + x^5 + 3x^6, \quad B(x) = 2x^2 + x^3 + x^4 + 2x^5.$$

Multiplying, each pair of faces contributes  $x^i x^j = x^{i+j}$ , so the coefficient of  $x^t$  in  $A(x)B(x)$  is the number of ways to throw a total of  $t$ . This is the generating function idea of Section 9.4 once more: the table of totals is just the product  $A(x)B(x)$ , and our case-by-case deduction was a way of factorising it.

### Sicherman's dice

For two ordinary dice the product is  $(x + x^2 + \dots + x^6)^2$ . Any pair of dice with the same totals must have polynomials  $P(x), Q(x)$  with  $P(x)Q(x) = (x + \dots + x^6)^2$ . Now factorise:

$$x + x^2 + \dots + x^6 = x(1 + x)(1 + x + x^2)(1 - x + x^2),$$

and these factors cannot be broken down further (into polynomials with integer coefficients). So each of  $P$  and  $Q$  must be built from the factors  $x$ ,  $1 + x$ ,  $1 + x + x^2$ ,  $1 - x + x^2$ , each appearing twice in total. Three conditions narrow the choice:

- each die has 6 faces, so  $P(1) = Q(1) = 6$ ; the factors take the values 1, 2, 3, 1 at  $x = 1$ , so each die gets one  $1 + x$  and one  $1 + x + x^2$ ;
- no face shows 0, so  $P(0) = Q(0) = 0$ : each die gets one factor  $x$ ;
- the only freedom left is how the two copies of  $1 - x + x^2$  are shared. One each gives ordinary dice. Both on one die gives

$$x(1+x)(1+x+x^2) = x + 2x^2 + 2x^3 + x^4,$$

$$x(1+x)(1+x+x^2)(1-x+x^2)^2 = x + x^3 + x^4 + x^5 + x^6 + x^8.$$

These are Sicherman's dice, 1, 2, 2, 3, 3, 4 and 1, 3, 4, 5, 6, 8 (Figure 9.6.1), and the argument proves they are the only alternative. The key fact is that polynomials, like integers, factorise uniquely into irreducibles: number theory and combinatorics meet.

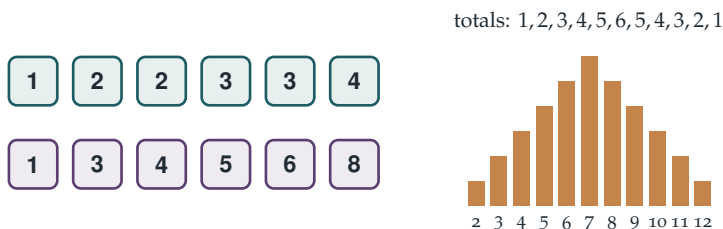


**Figure 9.6.1** Sicherman's dice. Thrown together, they give every total from 2 to 12 with exactly the same frequency as two ordinary dice (right): one way to throw 2, two ways to throw 3, and so on up to six ways to throw 7.

### Probability generating functions

In probability the same idea appears as the *probability generating function*: a random variable  $X$  taking values  $0, 1, 2, \dots$  has  $G(t) = E(t^X) = \sum P(X = k)t^k$ , and for independent  $X$  and  $Y$ ,  $E(t^{X+Y}) = E(t^X)E(t^Y)$ . Sums of independent random variables correspond to products of generating functions. STEP III 2003 Q14 (in the exercises) finds Sicherman's dice this way, and Section 9.13 uses a related argument to show that no loading of two dice can make all totals equally likely.

## The thinking skill

### Encode a counting problem as a polynomial

*When objects combine by adding their values, represent each one as a polynomial in  $x$ . Combining becomes multiplying, and questions about the combinations become questions about factorisation.*

Here a question about dice turned into the unique factorisation of polynomials. At university, such encodings are everywhere: in combinatorics, in probability, and in algebra, where polynomials are studied in their own right as a ring with arithmetic like that of the integers.

## Try this next

- 1 Tetrahedral dice.** (From STEP III, 2003.) Standard four-faced dice are labelled 1, 2, 3, 4. Find all the ways in which two four-faced dice can be labelled with positive integers so that every total has the same probability as for two standard four-faced dice. (*Hint:*  $x + x^2 + x^3 + x^4 = x(1 + x)(1 + x^2)$ .)
- 2 Checking the table.** Multiply out  $(x + x^4 + x^5 + 3x^6)(2x^2 + x^3 + x^4 + 2x^5)$  and check it against the table in the question.
- 3 Three dice.** Are there three six-sided dice, not all standard, labelled with positive integers, whose totals have the same distribution as three standard dice? (*Hint:* the factors of  $(x + \dots + x^6)^3$  now include three copies of  $1 - x + x^2$ .)
- 4 A coin and an octahedron.** An eight-sided die labelled 1, ..., 8 and a two-sided “die” (a coin) labelled 1, 2: can they be relabelled differently with the same totals?

## Further reading

- **Martin Gardner, *Penrose Tiles to Trapdoor Ciphers*** (Mathematical Association of America). Includes the column on Sicherman dice.
- **Joseph Gallian & David Rusin, “Cyclotomic polynomials and nonstandard dice”, *Discrete Mathematics* 27 (1979).** The general theory, for dice with any number of faces.
- **Herbert Wilf, *generatingfunctionology*** (free online), Chapter 1.



# 9.7 Equal Sums of Powers

## The question

Let  $p_0(x) = (1-x)(1-x^2)(1-x^4)$ . Show that  $(1-x)^3$  is a factor of  $p_0(x)$ . If  $p_1(x) = xp'_0(x)$  show, by considering factors of the polynomials involved, that  $p'_0(1) = 0$  and  $p'_1(1) = 0$ .

By writing  $p_0(x)$  in the form

$$p_0(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + c_4x^4 + c_5x^5 + c_6x^6 + c_7x^7,$$

deduce that

$$\begin{aligned} 1 + 2 + 4 + 7 &= 3 + 5 + 6 \\ 1^2 + 2^2 + 4^2 + 7^2 &= 3^2 + 5^2 + 6^2. \end{aligned}$$

Show that we can write the integers  $1, 2, \dots, 15$  in some order as  $a_1, a_2, \dots, a_{15}$  in such a way that

$$a_1^r + a_2^r + \dots + a_8^r = a_9^r + a_{10}^r + \dots + a_{15}^r$$

for  $r = 1, 2, 3$ .

## The hook

Split the numbers 1 to 16 into two teams of eight:

$$\{1, 4, 6, 7, 10, 11, 13, 16\} \quad \text{and} \quad \{2, 3, 5, 8, 9, 12, 14, 15\}.$$

Each team sums to 68. The sums of their squares are both 748. The sums of their cubes are both 9248. It is hard to believe this could be anything but a fluke, yet it is the third step of a pattern that goes on for ever: split 1 to 32 the right way and the fourth powers agree as well.

The split was found by Eugène Prouhet in 1851, and the rule behind it, the *Thue–Morse sequence*, turns up in an astonishing range of places: in number theory, in the theory of computation, in chess, and in the fairest way to take turns. This question uncovers it through a polynomial and its derivatives.

## Before you start

**?** **Ask yourself** If a polynomial  $p(x) = \sum c_k x^k$  has  $p(1) = 0$ , what does that say about the coefficients? What about  $p'(1) = 0$ ?

$p(1) = \sum c_k$ , so  $p(1) = 0$  says the coefficients add to zero. And  $p'(x) = \sum k c_k x^{k-1}$ , so  $p'(1) = \sum k c_k$ . If the coefficients are all  $\pm 1$ , these say that the positions of the  $+1$ s and the  $-1$ s are equal in number, and have equal sums.

## Working it through

### STAGE 1 A triple root

**?** **Ask yourself** Factorise  $1 - x^2$  and  $1 - x^4$ . How many factors of  $1 - x$  appear?

Since  $1 - x^2 = (1 - x)(1 + x)$  and  $1 - x^4 = (1 - x)(1 + x)(1 + x^2)$ ,

$$p_0(x) = (1 - x)^3(1 + x)^2(1 + x^2),$$

so  $(1 - x)^3$  is a factor. ■

### STAGE 2 Derivatives at 1

**?** **Ask yourself** If  $(1 - x)^3$  divides  $p_0$ , what power of  $(1 - x)$  divides  $p'_0$ ? And then  $p_1 = x p'_0$  and  $p'_1$ ?

Write  $p_0(x) = (1 - x)^3 q(x)$ . By the product rule,  $p'_0(x) = -3(1 - x)^2 q(x) + (1 - x)^3 q'(x)$ , which has  $(1 - x)^2$  as a factor. So  $p'_0(1) = 0$ . Then  $p_1(x) = x p'_0(x)$  also has  $(1 - x)^2$  as a factor, and the same argument shows  $p'_1$  has  $(1 - x)$  as a factor, so  $p'_1(1) = 0$ . ■

### STAGE 3 Reading off the coefficients

Multiplying out,  $(1 - x)(1 - x^2) = 1 - x - x^2 + x^3$ , and then

$$p_0(x) = 1 - x - x^2 + x^3 - x^4 + x^5 + x^6 - x^7.$$

**?** **Ask yourself** What are  $p'_0(1)$  and  $p'_1(1)$  in terms of the coefficients  $c_k$ ?

Since  $p_1(x) = x p'_0(x) = \sum k c_k x^k$ , we have  $p'_1(1) = \sum k^2 c_k$ . The coefficient is  $+1$  at  $k = 0, 3, 5, 6$  and  $-1$  at  $k = 1, 2, 4, 7$ . So  $p'_0(1) = 0$  and  $p'_1(1) = 0$  say

$$0 + 3 + 5 + 6 = 1 + 2 + 4 + 7 \quad (= 14),$$

$$0^2 + 3^2 + 5^2 + 6^2 = 1^2 + 2^2 + 4^2 + 7^2 \quad (= 70),$$

which, dropping the 0, are the identities required. ■

#### STAGE 4 One more factor



**Pause and try** To make the cubes balance too, we need a fourth factor of  $(1 - x)$ . Which polynomial should replace  $p_0$ ?

Take  $P(x) = (1 - x)(1 - x^2)(1 - x^4)(1 - x^8)$ , which has  $(1 - x)^4$  as a factor. Its coefficients are  $p_0(x) - x^8 p_0(x)$ : the pattern of signs for 0 to 7, followed by the same pattern reversed in sign for 8 to 15:

|       |   |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |
|-------|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| $k$   | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| $c_k$ | + | - | - | + | - | + | + | - | - | + | +  | -  | +  | -  | -  | +  |

The argument of Stage 2, applied one more time, shows that  $\sum c_k k^r = 0$  for  $r = 0, 1, 2, 3$ . Take the  $+$  positions  $\{0, 3, 5, 6, 9, 10, 12, 15\}$  and the  $-$  positions  $\{1, 2, 4, 7, 8, 11, 13, 14\}$ . For  $r = 1, 2, 3$ , the term from 0 is zero, so it can be dropped. So the order

$$a_1, \dots, a_8 = 1, 2, 4, 7, 8, 11, 13, 14, \quad a_9, \dots, a_{15} = 3, 5, 6, 9, 10, 12, 15$$

works: both halves sum to 60, their squares to 620, and their cubes to 7200. ■

### Looking back

- **Derivatives detect repeated roots.** A factor  $(1 - x)^m$  forces the first  $m - 1$  derivatives to vanish at  $x = 1$ .
- **Multiplying by  $x$  before differentiating** turned  $\sum k c_k$  into  $\sum k^2 c_k$ , and so on to higher powers.
- **Doubling.** Each extra factor  $(1 - x^{2^m})$  doubled the length and added one more power that balances.

### What's really going on

#### The Thue–Morse sequence

The signs of the coefficients of  $(1 - x)(1 - x^2)(1 - x^4)(1 - x^8) \cdots$  form the *Thue–Morse sequence*. Writing 0 for  $+$  and 1 for  $-$ :

0110100110010110 ...

There are three equivalent descriptions.

- **Doubling:** start with 0, and repeatedly append the complement (swap 0s and 1s) of what you have:  $0 \rightarrow 01 \rightarrow 0110 \rightarrow 01101001 \rightarrow \cdots$ .

- **Binary digits:** the  $k$ th term is 0 if  $k$  has an even number of 1s when written in binary, and 1 if odd. (This is why the sign of  $x^k$  is  $(-1)^{\text{number of 1s in } k}$ : to make  $x^k$  from the product, you choose the factors  $x^{2^m}$  that make up  $k$  in binary, and each brings a minus sign.)
- **Polynomial:** the coefficients of  $\prod_m (1 - x^{2^m})$ , as in the question.

Splitting  $0, 1, \dots, 2^N - 1$  by the Thue–Morse sequence gives two sets whose  $r$ th powers have equal sums for every  $r < N$  (Figure 9.7.1). This is the *Prouhet–Tarry–Escott problem*: find two different sets of integers with equal sums of  $r$ th powers for  $r = 1, \dots, k$ . Prouhet’s solution uses  $2^{k+1}$  numbers; it is a famous open problem to find out how few numbers are really needed.

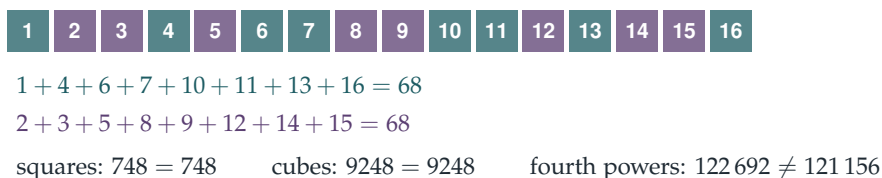


**Figure 9.7.1** The numbers 1 to 16 split by the Thue–Morse sequence. The two sets have equal sums, equal sums of squares and equal sums of cubes, but not equal sums of fourth powers; for that you need 1 to 32. Reading the colours left to right gives the pattern teal–plum–plum–teal, plum–teal–teal–plum, ...

### The fairest way to take turns

Two children take turns choosing sweets from a pile, best first. If they alternate  $ABABAB\dots$ , then  $A$  always gets the better sweet of each pair, a systematic advantage. The order  $ABBA$  is fairer, and  $ABBABAAB$  fairer still: this is the Thue–Morse sequence, and the balanced power sums are exactly the reason. If the sweets have values given by a smooth function of their position, expanding that function as a polynomial shows that the Thue–Morse order cancels its leading terms. The same idea has been proposed for penalty shoot-outs in football (the “ $ABBA$ ” order trialled by football’s rule-makers from 2017) and is used for the order of serving in tennis tie-breaks.

### Thue, Morse and Euwe

The sequence has been discovered many times. Prouhet used it in 1851. Axel Thue rediscovered it in 1912 while studying sequences without repetitive patterns: it never contains the same block three times in a row (it is “cube-free”). Marston Morse found it again in 1921 in the study of geodesics on surfaces. And the chess world champion Max Euwe used it in 1929 to show that the rule then in force for drawing a game by repetition did not prevent infinitely long games. Today it is a central example in the theory of *automatic sequences*, sequences computed by simple machines reading the binary digits of  $n$ .

## The thinking skill

### Encode a combinatorial property in a polynomial

*Balanced sums of powers are about where the  $+$  and  $-$  signs fall. A polynomial with those signs as coefficients turns the property into a statement about roots.*

The question never asks you to search: a repeated root at  $x = 1$  does all the work. Using algebra to produce combinatorial objects, rather than finding them by trial, is a recurring idea in university combinatorics.

## Try this next

- 1 **Shifting.** (From STEP II, 2025.) Suppose  $\sum a_k = \sum b_k$  and  $\sum a_k^2 = \sum b_k^2$  for two lists of the same length  $m$ . Show that  $\sum a_k^3 + \sum (c + b_k)^3 = \sum b_k^3 + \sum (c + a_k)^3$  for every  $c$ . Use this, starting from a balanced split of  $\{1, \dots, 8\}$ , to find a split of  $\{1, \dots, 16\}$  in which the cubes balance as well.
- 2 **Squares of squares.** (Also from STEP II, 2025.) The sets  $\{1, 3, 4, 5, 9, 11\}$  and  $\{2, 6, 7, 8, 10\}$  have equal sums and equal sums of squares. Split  $\{n^2, (n+1)^2, \dots, (n+11)^2\}$  into two sets with equal sums, for every positive integer  $n$ . (*Hint:* add 0 to the second set.)
- 3 **Translation invariance.** Show that if two sets of numbers have equal sums of  $r$ th powers for  $r = 0, 1, \dots, k$ , then so do the sets obtained by adding the same number  $c$  to every element. Why does this mean that the question's sets  $\{1, \dots, 15\}$  could equally be replaced by any 16 consecutive integers?
- 4 **Fair sweets.** Sweets have values  $1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{8}$ . Compare the totals  $A$  gets with the orders  $ABABABAB$  and  $ABBABAAB$ .

## Further reading

- **Jean-Paul Allouche & Jeffrey Shallit**, “The ubiquitous Prouhet–Thue–Morse sequence”, in *Sequences and Their Applications* (Springer, 1999). A survey of the sequence's many appearances; free online.
- **Matt Parker**, *Things to Make and Do in the Fourth Dimension* (Penguin). Includes equal power sums and other number curiosities.
- **Joshua Cooper & Aaron Dutle**, “Greedy Galois games”, *American Mathematical Monthly* 120 (2013). Fair turn-taking and Thue–Morse.



## 9.8 Nobody Gets Their Own Shirt

### The question

Starting with the result  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ , prove that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C).$$

Write down, without proof, the corresponding result for four events  $A, B, C$  and  $D$ .

A pack of  $n$  cards, numbered  $1, 2, \dots, n$ , is shuffled and laid out in a row. The result of the shuffle is that each card is equally likely to be in any position in the row. Let  $E_i$  be the event that the card bearing the number  $i$  is in the  $i$ th position in the row. Write down the following probabilities:

- (i)  $P(E_i)$ ;
- (ii)  $P(E_i \cap E_j)$ , where  $i \neq j$ ;
- (iii)  $P(E_i \cap E_j \cap E_k)$ , where  $i \neq j, j \neq k$  and  $k \neq i$ .

Hence show that the probability that at least one card is in the same position as the number it bears is

$$1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n+1} \frac{1}{n!}.$$

Find the probability that exactly one card is in the same position as the number it bears.

### The hook

A football team of eleven players changes in a hurry and grabs shirts at random. What is the chance that nobody ends up in their own shirt? Now suppose it is a squad of fifty, or a crowd of a million. Does the chance go up, because there are more people to get it wrong? Or down, because there are more people who might get it right by luck? Remarkably, it hardly changes at all: from five people onwards, the chance is within half a per cent of  $1/e \approx 0.368$ .

This is the oldest problem in combinatorial probability, posed by Pierre de Montmort in 1708 as the *problème des rencontres*. The question solves it with the *inclusion–exclusion principle*, a tool for counting things that overlap, and the answer is one of the most unexpected appearances of  $e$  in mathematics.

## Before you start



**Pause and try** Try  $n = 3$ : list all six orders of cards 1, 2, 3 and count how many have no card in its own place. Then try  $n = 4$ , if you are patient (there are 24 orders).

For  $n = 3$  only 231 and 312 have no card in its place, so the probability is  $\frac{2}{6} = \frac{1}{3}$ . For  $n = 4$  there are 9 such orders out of 24, giving  $\frac{3}{8} = 0.375$ . An arrangement with no fixed card is called a *derangement*.

## Working it through

### STAGE 1 Inclusion–exclusion for three events



**Ask yourself** How can you write  $A \cup B \cup C$  as a union of two events, so that the two-event formula applies?

Treat  $A \cup B$  as a single event and apply the formula to it and  $C$ :

$$P(A \cup B \cup C) = P(A \cup B) + P(C) - P((A \cup B) \cap C).$$

Now  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ , and applying the formula once more,  $P((A \cap C) \cup (B \cap C)) = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)$ . Expanding  $P(A \cup B)$  too and collecting terms gives the result. ■

For four events the pattern continues: add the four single probabilities, subtract the six pairwise intersections, add the four triple intersections, and subtract  $P(A \cap B \cap C \cap D)$ .

### STAGE 2 The probabilities



**Ask yourself** If card  $i$  is in position  $i$ , how many ways are there to arrange the rest? And if cards  $i$  and  $j$  are both fixed?

There are  $n!$  equally likely orders. If card  $i$  is fixed in position  $i$ , the other  $n - 1$  cards can be in any order:  $(n - 1)!$  ways. So

$$P(E_i) = \frac{(n-1)!}{n!} = \frac{1}{n}, \quad P(E_i \cap E_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}, \quad P(E_i \cap E_j \cap E_k) = \frac{1}{n(n-1)(n-2)}.$$

In general, the probability that  $k$  specified cards are all in their own positions is  $\frac{(n-k)!}{n!}$ .

### STAGE 3 Adding it up

**Ask yourself** In the inclusion–exclusion formula for  $P(E_1 \cup \cdots \cup E_n)$ , how many terms are there involving intersections of  $k$  events, and what is their total?

There are  $\binom{n}{k}$  ways to choose  $k$  of the events, and each intersection has probability  $\frac{(n-k)!}{n!}$ . So the  $k$ -fold terms add up to

$$\binom{n}{k} \frac{(n-k)!}{n!} = \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{n!} = \frac{1}{k!}.$$

Everything except the  $k!$  cancels. With alternating signs,

$$P(\text{at least one card in its place}) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1} \frac{1}{n!}. \quad \blacksquare$$

So the probability of a derangement is

$$d_n = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} = \sum_{k=0}^n \frac{(-1)^k}{k!}.$$

#### STAGE 4 Exactly one

**Ask yourself** For exactly one card to be in place, which card is it, and what must the others do?

Choose which card is in its place ( $n$  ways). The probability that card  $i$  is in place is  $\frac{1}{n}$ , and given that, the other  $n-1$  cards must form a derangement among themselves, which has probability  $d_{n-1}$ . So

$$P(\text{exactly one}) = n \cdot \frac{1}{n} \cdot d_{n-1} = \sum_{k=0}^{n-1} \frac{(-1)^k}{k!}. \quad \blacksquare$$

#### Looking back

- **Overcount, then correct.** Adding the  $P(E_i)$  counts arrangements with two fixed cards twice; subtracting the pairs corrects this but goes too far for triples; and so on.
- **Symmetry makes the terms equal.** Every  $k$ -fold intersection has the same probability, so each layer of the formula is one number times a binomial coefficient.
- **Reduce to a smaller case.** “Exactly one” became a derangement of  $n-1$  cards.

#### What’s really going on

##### Where $e$ comes from

The series  $\sum (-1)^k / k!$  is the exponential series  $e^x = \sum x^k / k!$  at  $x = -1$ . So as  $n \rightarrow \infty$ ,

$$d_n \rightarrow e^{-1} \approx 0.3679,$$

and the probability of exactly one match tends to the same limit. The convergence is extraordinarily fast (Figure 9.8.1): the error in stopping at  $1/n!$  is less than  $1/(n+1)!$ , so already for  $n = 7$  the answer agrees with  $1/e$  to four decimal places. In fact the number of derangements of  $n$  objects,  $n!d_n$ , is exactly the nearest integer to  $n!/e$ .

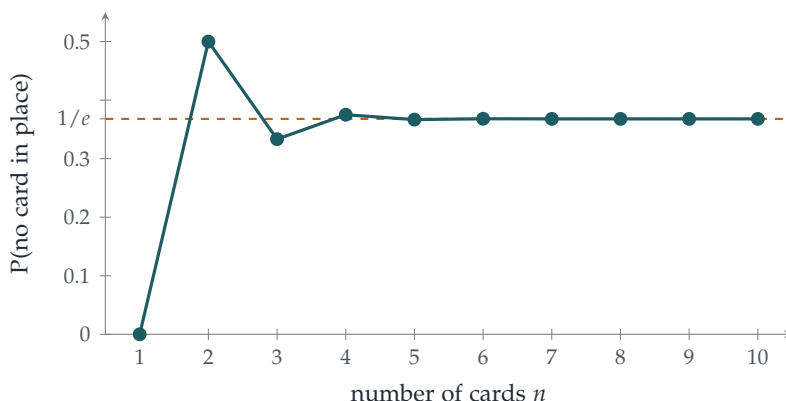


**Figure 9.8.1** The probability that a random shuffle of  $n$  cards leaves no card in its own place. It alternates above and below  $1/e$ , and from  $n = 6$  onwards it is indistinguishable from  $1/e$  on this scale.

### The number of matches is Poisson

More is true. Let  $N$  be the number of cards in their own place. Then  $P(N = k) \rightarrow \frac{e^{-1}}{k!}$  for each  $k$ : the number of matches has, in the limit, the *Poisson distribution* with mean 1. You can see why the mean is 1 without any calculation:  $N$  is the sum of the indicators of  $E_1, \dots, E_n$ , each with expectation  $\frac{1}{n}$ , and expectation adds even though the events are not independent. STEP III 2004 Q12 (in the exercises) shows that the variance is 1 too. The Poisson distribution appears whenever there are many unlikely events that are nearly independent; derangements are a classic example, and the subject of Section 12.11 in Chapter 12.

### Montmort's game of *treize*

Pierre Rémond de Montmort, a French nobleman, studied games of chance in his *Essay d'analyse sur les jeux de hazard* (1708). In the game of *treize* (thirteen), a player deals thirteen cards, calling out “one, two, three, ...” and wins if the card dealt ever matches the number called. Montmort found the probability of winning, and in correspondence with Nicolaus Bernoulli the general formula was established by 1713. Euler rediscovered it in 1753. The inclusion–exclusion principle was later formalised by Abraham de Moivre and, in its general form, by James Joseph Sylvester in the nineteenth century.

## The thinking skill

### Inclusion–exclusion

*To count a union of overlapping sets, add the sizes of the sets, subtract the pairwise overlaps, add the triple overlaps, and so on.*

It is most powerful when symmetry makes all the overlaps of each size equal, as here. At university it appears in combinatorics, in probability and in number theory, where it counts integers not divisible by any of a list of primes, and leads to Euler's totient function and the Möbius function.

## Try this next

- 1 **Shirts.** (From STEP III, 2004.) A team of  $m$  players put on  $m$  numbered shirts at random. Let  $C_i = 1$  if player  $i$  wears shirt  $i$ , and 0 otherwise. Find  $E(C_1)$ ,  $\text{Var}(C_1)$  and  $\text{Cov}(C_1, C_2)$ , and show that  $N = C_1 + \cdots + C_m$  has  $E(N) = \text{Var}(N) = 1$ . For  $m = 4$ , compare  $P(N = 0)$  with the Poisson approximation  $e^{-1}$ .
- 2 **A recurrence.** Let  $D_n$  be the number of derangements of  $n$  objects. By considering where card 1 goes, and whether the card in that position goes to position 1, show that  $D_n = (n - 1)(D_{n-1} + D_{n-2})$ . Check it against  $D_3 = 2$ ,  $D_4 = 9$ .
- 3 **Exactly  $k$ .** Show that the probability that exactly  $k$  cards are in their own places is  $\frac{1}{k!}d_{n-k}$ , and that it tends to  $\frac{e^{-1}}{k!}$ .
- 4 **Coprime integers.** Use inclusion–exclusion to count the integers from 1 to 30 divisible by none of 2, 3 and 5. How does your answer relate to  $30(1 - \frac{1}{2})(1 - \frac{1}{3})(1 - \frac{1}{5})$ ?

## Further reading

- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press). A first-year text that covers inclusion–exclusion and the matching problem.
- **William Feller, *An Introduction to Probability Theory and Its Applications, Vol. 1*** (Wiley), Chapter IV. A classic, with many variations on the matching problem.
- **Miklós Bóna, *A Walk Through Combinatorics*** (World Scientific), the chapter on inclusion–exclusion.



## 9.9 The Birthday Problem

### The question

Given that  $0 < r < n$  and  $r$  is much smaller than  $n$ , show that  $\frac{n-r}{n} \approx e^{-r/n}$ .

There are  $k$  guests at a party. Assuming that there are exactly 365 days in the year, and that the birthday of any guest is equally likely to fall on any of these days, show that the probability that there are at least two guests with the same birthday is approximately  $1 - e^{-k(k-1)/730}$ .

Using the approximation  $\frac{253}{365} \approx \ln 2$ , find the smallest value of  $k$  such that the probability that at least two guests share the same birthday is at least  $\frac{1}{2}$ .

How many guests must there be at the party for the probability that at least one guest has the same birthday as the host to be at least  $\frac{1}{2}$ ?

### The hook

How many people do you need in a room for it to be more likely than not that two of them share a birthday? Most people guess somewhere around 180, half of 365. The answer is 23. In a class of thirty there is a 70% chance of a shared birthday; at a party of fifty, 97%.

The surprise comes from confusing two different questions: “does someone share *my* birthday?” (which does need about 253 people) and “do *any two people* share a birthday?”. With 23 people there are  $\binom{23}{2} = 253$  pairs, and each pair is a chance for a match. This question makes that intuition precise, and the same mathematics decides how secure the cryptography protecting your bank account is.

### Before you start

**Ask yourself** Why is it easier to calculate the probability that *no* two guests share a birthday? Line the guests up: what is the chance that the second avoids the first’s birthday? That the third avoids both?

“At least two share” covers many overlapping cases (two share, three share, two different pairs, ...). Its complement, “all different”, is a single clean event. The second guest avoids the first’s birthday with probability  $\frac{364}{365}$ ; given that, the third avoids both with probability  $\frac{363}{365}$ ; and so on.

## Working it through

### STAGE 1 A useful approximation

**?** **Ask yourself** What is the tangent line to  $y = e^{-x}$  at  $x = 0$ ?

Since  $e^{-x} = 1 - x + \frac{x^2}{2} - \dots$ , for small  $x$  we have  $e^{-x} \approx 1 - x$ . With  $x = r/n$ ,

$$\frac{n-r}{n} = 1 - \frac{r}{n} \approx e^{-r/n}. \quad \blacksquare$$

The error is about  $\frac{r^2}{2n^2}$ , small when  $r$  is much smaller than  $n$ .

### STAGE 2 The probability of a shared birthday

The probability that all  $k$  birthdays are different is

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - (k-1)}{365} = \prod_{r=0}^{k-1} \frac{365 - r}{365}.$$

Using the approximation on each factor, and the fact that exponentials multiply by adding their exponents,

$$\prod_{r=0}^{k-1} e^{-r/365} = \exp\left(-\frac{0+1+\cdots+(k-1)}{365}\right) = \exp\left(-\frac{k(k-1)}{730}\right).$$

So the probability of at least one shared birthday is approximately  $1 - e^{-k(k-1)/730}$ .  $\blacksquare$

**?** **Ask yourself** Look at the exponent:  $\frac{k(k-1)}{730} = \binom{k}{2} / 365$ . What does that say in words?

It is the number of pairs of guests, times the probability  $\frac{1}{365}$  that a given pair shares a birthday: the expected number of shared pairs. The approximation says the chance of *no* match is  $e^{-(\text{expected number of matches})}$ , a hint of the Poisson distribution again, as in Section 9.8.

### STAGE 3 Twenty-three

We need  $1 - e^{-k(k-1)/730} \geq \frac{1}{2}$ , that is  $e^{-k(k-1)/730} \leq \frac{1}{2}$ , or

$$\frac{k(k-1)}{730} \geq \ln 2 \approx \frac{253}{365}, \quad \text{that is} \quad k(k-1) \geq 506.$$

Since  $22 \times 21 = 462$  and  $23 \times 22 = 506$ , the smallest value is  $k = 23$ .  $\blacksquare$

(The exact probability for 23 guests is 0.507, and for 22 guests 0.476, so the approximation gives the right answer.)

### STAGE 4 Sharing with the host

**Ask yourself** For each guest, what is the probability of *not* sharing the host's birthday? Are the guests independent of each other in this respect?

Each guest independently misses the host's birthday with probability  $\frac{364}{365} \approx e^{-1/365}$ . So the probability that none of  $k$  guests shares it is  $(\frac{364}{365})^k \approx e^{-k/365}$ , and we need

$$e^{-k/365} \leq \frac{1}{2}, \quad \text{that is} \quad k \geq 365 \ln 2 \approx 253.$$

So there must be 253 guests. ■

The contrast is the whole point: 253 guests to match one particular person, but only 23 for some pair to match, because 23 people make 253 pairs.

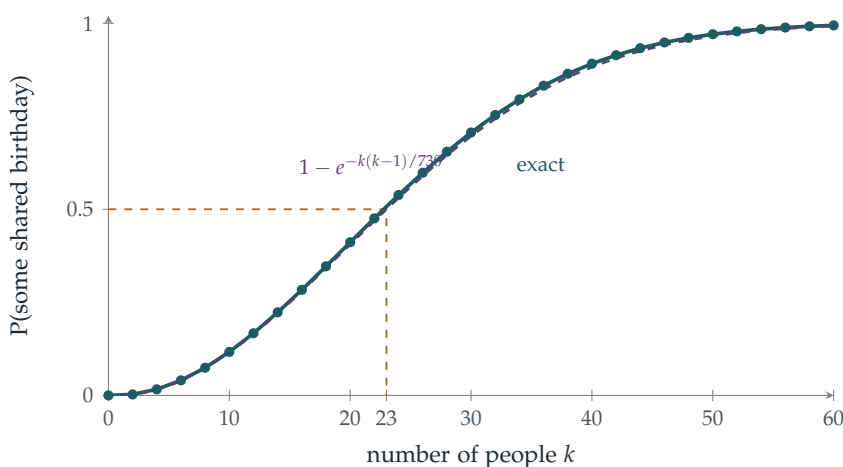


**Figure 9.9.1** The probability that at least two of  $k$  people share a birthday: exact values (teal) and the approximation  $1 - e^{-k(k-1)/730}$  (plum, dashed). The two are very close, and both pass  $\frac{1}{2}$  between  $k = 22$  and  $k = 23$ .

## Looking back

- **Use the complement.** “All different” was a single product; “at least one match” was a mess.
- **Linearise, then exponentiate.** Replacing each factor  $1 - x$  by  $e^{-x}$  turned a product into a sum.
- **Count pairs.** The exponent was the number of pairs times the chance that a pair matches. The square in  $k(k-1)$  is the whole explanation of the paradox.

## What's really going on

### The square-root law

With  $n$  equally likely days instead of 365, the same argument gives

$$P(\text{some match among } k) \approx 1 - e^{-k^2/(2n)},$$

so matches become likely once  $k$  is around  $\sqrt{n}$ , not  $n$ . More precisely, the probability reaches  $\frac{1}{2}$  at  $k \approx 1.18\sqrt{n}$ , and the expected number of people you need to meet before the first repeat is about  $\sqrt{\pi n/2}$  (for  $n = 365$ , about 24.6). This *square-root law* is the reason coincidences are so common: in any large collection of events, some pair is likely to match in some way, even though any particular match is unlikely.

### Birthday attacks

Computers check that a file has not been tampered with by computing a *hash*: a fingerprint of, say,  $n = 2^{128}$  possible values. A forger who wants two documents with the same hash (a harmless contract and a fraudulent one) does not need to try  $2^{128}$  documents. By the birthday argument, generating about  $\sqrt{2^{128}} = 2^{64}$  variants of each will probably produce a match. This *birthday attack* is why modern hash functions use 256 bits or more: half the bits are lost to the square-root law. Similar calculations decide how long random identifiers must be so that two computers never accidentally choose the same one.

#### Who first asked?

The birthday problem is usually credited to Richard von Mises, who discussed it in 1939, though it circulated earlier. Harold Davenport is also said to have known it earlier, and not to have published it because it seemed too simple. As Warren Weaver told the story, he liked to demonstrate it at dinners of army officers during the Second World War: he bet that two of the twenty-two people present shared a birthday, and lost, until a waitress revealed that she shared her birthday with one of the officers.

### The thinking skill

#### Calculate the complement; approximate carefully

When “at least one” is complicated, find the probability of “none”. When a product is awkward, approximate each factor by an exponential and add the exponents.

Both moves are standard at university. The second,  $1 - x \approx e^{-x}$ , is the start of the theory of Poisson approximation and of the asymptotic methods used to estimate quantities too complicated to calculate exactly.

## Try this next

---

- 1 The first repeat.** (From STEP II, 2014.) A random number generator produces integers from 1 to  $n$ , independently and equally likely. Let  $X$  be the position of the first repeat. Find  $P(X = r)$  for  $2 \leq r \leq n + 1$ , and deduce that

$$\frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{2}{n} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\frac{3}{n} + \cdots = 1.$$

Show that  $E(Y) = \sum_{k=1}^N P(Y \geq k)$  for any random variable taking values  $1, \dots, N$ , and hence that

$$\left(1 - \frac{1^2}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2^2}{n}\right) + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3^2}{n}\right) + \cdots = 0.$$

- 2 Three in common.** Roughly how many people are needed for a better-than-even chance that some *three* share a birthday? (*Hint:* the expected number of triples sharing a birthday is  $\binom{k}{3}/365^2$ . Set this to about  $\ln 2$ .) The true answer is 88.
- 3 Near misses.** How many people are needed for an even chance that two have birthdays on the same day or on adjacent days?
- 4 Uneven birthdays.** Real birthdays are not uniform. Show that if the days have probabilities  $p_1, \dots, p_{365}$ , the chance that two people share a birthday is  $\sum p_i^2$ , and that this is smallest when all  $p_i$  are equal. (So non-uniformity only makes matches more likely.)

## Further reading

---

- **William Feller**, *An Introduction to Probability Theory and Its Applications*, Vol. 1 (Wiley), Chapter II. The birthday problem and its approximations.
- **Persi Diaconis & Frederick Mosteller**, “Methods for studying coincidences”, *Journal of the American Statistical Association* 84 (1989). Why coincidences are common; very readable.
- **David Hand**, *The Improbability Principle* (Bantam). The square-root law and other reasons why unlikely things happen all the time.



# 9.10 Always Able to Give Change

## The question

I am selling raffle tickets for £1 per ticket. In the queue for tickets, there are  $m$  people each with a single £1 coin and  $n$  people each with a single £2 coin. Each person in the queue wants to buy a single raffle ticket and each arrangement of people in the queue is equally likely to occur. Initially, I have no coins and a large supply of tickets. I stop selling tickets if I cannot give the required change.

- (i) In the case  $n = 1$  and  $m \geq 1$ , find the probability that I am able to sell one ticket to each person in the queue.
- (ii) By considering the first three people in the queue, show that the probability that I am able to sell one ticket to each person in the queue in the case  $n = 2$  and  $m \geq 2$  is  $\frac{m-1}{m+1}$ .
- (iii) Show that the probability that I am able to sell one ticket to each person in the queue in the case  $n = 3$  and  $m \geq 3$  is  $\frac{m-2}{m+1}$ .

## The hook

The pattern is irresistible:  $\frac{m}{m+1}, \frac{m-1}{m+1}, \frac{m-2}{m+1}$ . Surely, with  $n$  people holding £2 coins, the answer is  $\frac{m-n+1}{m+1}$ . It is, and the slickest proof is one of the most elegant ideas in combinatorics: the *reflection principle*. It turns a hard count (queues that never run out of change) into an easy one by reflecting part of a picture in a line.

The same problem has many disguises: an election count in which the winner is never behind (Bertrand's ballot problem of 1887), a gambler who never goes broke, brackets that match correctly. When  $m = n$  the count gives the *Catalan numbers*, perhaps the most frequently occurring sequence in combinatorics after the binomial coefficients.

## Before you start



**Ask yourself** Each £1 customer gives you a coin; each £2 customer takes one away (as change). When do you get stuck?

Keep track of the number of £1 coins you hold. It goes up by 1 for each £1 customer and down by 1 for each £2 customer. You are stuck exactly when a £2 customer arrives

and you hold no coins: the count would go to  $-1$ . So you succeed if and only if, at every point in the queue, the number of £1 customers so far is at least the number of £2 customers so far.

## Working it through

### STAGE 1 One £2 coin

With  $n = 1$ , you fail only if the £2 customer is first. The £2 customer is equally likely to be in any of the  $m + 1$  positions, so

$$P(\text{success}) = 1 - \frac{1}{m+1} = \frac{m}{m+1}. \quad \blacksquare$$

### STAGE 2 Two £2 coins

**Ask yourself** The first person must have a £1 coin. What can happen with the second person, and in each case, what situation are you left in?

The first person has a £1 coin with probability  $\frac{m}{m+2}$ ; otherwise you fail at once. Then you hold one coin, and the remaining queue has  $m - 1$  £1 people and two £2 people.

- ▶ If the second person has a £1 coin (probability  $\frac{m-1}{m+1}$ ), you hold two coins, enough to change both £2 coins whenever they come. Success is certain.
- ▶ If the second person has a £2 coin (probability  $\frac{2}{m+1}$ ), you are back to no coins, with  $m - 1$  £1 people and one £2 person left. By part (i) with  $m - 1$  in place of  $m$ , you succeed with probability  $\frac{m-1}{m}$ .

So

$$P(\text{success}) = \frac{m}{m+2} \left( \frac{m-1}{m+1} + \frac{2}{m+1} \cdot \frac{m-1}{m} \right) = \frac{m}{m+2} \cdot \frac{m-1}{m+1} \cdot \frac{m+2}{m} = \frac{m-1}{m+1}. \quad \blacksquare$$

**Pause and try** Try the same approach for  $n = 3$ . It works, but the cases multiply. Is there a way to count the successful queues all at once?

### STAGE 3 Queues as paths

**Ask yourself** Draw the number of coins you hold against the number of people served. What does a queue look like? What does a successful queue look like?

A queue is a path made of  $m$  up-steps (£1) and  $n$  down-steps (£2), starting at height 0 and ending at height  $m - n$ . There are  $\binom{m+n}{n}$  such paths, all equally likely. A queue is

successful exactly when its path never goes below 0. So we need to count the *bad* paths, those that touch height  $-1$ .

#### STAGE 4 The reflection

**Ask yourself** Take a bad path. Find the first moment it touches  $-1$ , and reflect the part of the path *before* that moment in the line  $y = -1$ . Where does the new path start, and where does it end?

The reflected path starts at  $-2$  (the mirror image of  $0$  in the line  $y = -1$ ) and ends where the original did, at  $m - n$  (Figure 9.10.1). Conversely, every path from  $-2$  to  $m - n$  must cross  $y = -1$  somewhere (it starts below and ends above, since  $m \geq n$ ); reflecting it before its first visit gives back a bad path from  $0$ . So bad paths from  $0$  correspond exactly to *all* paths from  $-2$  to  $m - n$ .

A path from  $-2$  to  $m - n$  in  $m + n$  steps rises by  $m - n + 2$ , so it has  $m + 1$  up-steps and  $n - 1$  down-steps: there are  $\binom{m+n}{n-1}$  of them. Hence

$$P(\text{failure}) = \frac{\binom{m+n}{n-1}}{\binom{m+n}{n}} = \frac{n! m!}{(n-1)! (m+1)!} = \frac{n}{m+1},$$

and

$$P(\text{success}) = 1 - \frac{n}{m+1} = \frac{m-n+1}{m+1}.$$

With  $n = 3$  this is  $\frac{m-2}{m+1}$ .

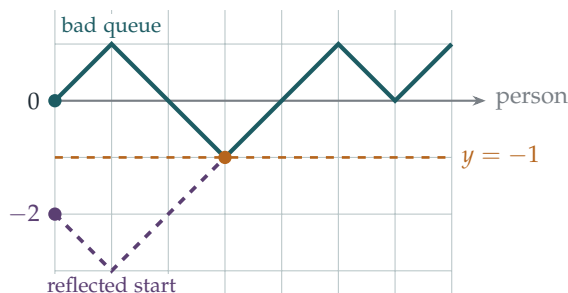


**Figure 9.10.1** A queue with  $m = 4$  £1 coins and  $n = 3$  £2 coins, drawn as a path (teal). It first hits  $-1$  at the third person (amber), when you cannot give change. Reflecting the path up to that point in the line  $y = -1$  gives a path from  $-2$  (plum, dashed) to the same end point. This matching between bad queues and paths starting at  $-2$  is the reflection principle.

### Looking back

- **Find the right state.** The only thing that mattered was the number of coins held, which turned the queue into a walk.
- **Condition on the first steps.** Parts (i) and (ii) were done by asking what the first few

people do, reducing to smaller cases.

- **Count the complement by a bijection.** The bad paths were hard to count directly, but reflection matched them with a set that was easy to count.

## What's really going on

---

### Catalan numbers

When  $m = n$  (exactly enough £1 coins), the number of successful queues is

$$\binom{2n}{n} - \binom{2n}{n-1} = \frac{1}{n+1} \binom{2n}{n},$$

the  $n$ th *Catalan number*  $C_n$ : 1, 1, 2, 5, 14, 42, 132, ... Richard Stanley's book *Catalan Numbers* lists more than two hundred different things that they count. Among them: the number of ways to arrange  $n$  pairs of brackets correctly (" $((()))()$ " is allowed, " $()()()$ " is not, by exactly the coin argument with "(" as £1 and ")" as £2), the number of ways to cut a polygon with  $n + 2$  sides into triangles by non-crossing diagonals, and the number of ways to bracket a product of  $n + 1$  factors.

### Random walks

Queues are random walks, and the reflection principle is one of the basic tools for studying them. It shows, for example, that a fair random walk starting at 0 is certain to return to 0 eventually, and that the probability that it stays strictly positive for the first  $2n$  steps is  $\frac{1}{2} \binom{2n}{n} 2^{-2n} \approx \frac{1}{2\sqrt{\pi n}}$ . In continuous form, the same reflection argument applied to *Brownian motion* gives the distribution of its maximum. This is the basis of the pricing of certain financial options, whose payoff depends on whether a share price ever crosses a barrier.

### Bertrand's ballot problem

In 1887 Joseph Bertrand asked: in an election where candidate  $A$  receives  $a$  votes and  $B$  receives  $b < a$ , what is the probability that  $A$  is strictly ahead throughout the count? The answer,  $\frac{a-b}{a+b}$ , was proved in the same year by Désiré André, whose argument is usually remembered as the reflection principle (although historians have pointed out that André's actual proof used a slightly different bijection). Our raffle problem is a variant in which ties are allowed: success means the £1 count never falls *below* the £2 count. One of the exercises deduces Bertrand's answer from ours.

---

## The thinking skill

### Bijjective counting

*To count a set, find a one-to-one correspondence with a set you can count.*

Reflection matched bad queues with unrestricted paths from a different start. Bijections are the most satisfying proofs in combinatorics, because they explain *why* two numbers are equal rather than just showing that they are. At university you will meet bijective proofs throughout combinatorics, and the reflection principle again in the study of random walks and Brownian motion.

## Try this next

- 1 **Three by cases.** Prove part (iii) by conditioning on the first people in the queue, as in part (ii), to see how much work the reflection principle saves.
- 2 **Bertrand's ballot.** In an election,  $A$  gets  $a$  votes and  $B$  gets  $b < a$ . Show that the probability that  $A$  is strictly ahead throughout the count is  $\frac{a-b}{a+b}$ . (*Hint:* the first vote must be for  $A$ ; after that,  $A$  must never fall behind  $B$  by more than. . . .)
- 3 **Brackets.** List the 5 correct arrangements of three pairs of brackets, and the 5 ways to cut a pentagon into triangles. Can you find a bijection between them?
- 4 **A £5 note.** A queue contains  $m$  people with £1 coins and  $n$  with £5 notes, for a ticket costing £1. Explain why the reflection principle no longer gives a simple answer. (This generalisation is much harder.)

## Further reading

- **William Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1** (Wiley), Chapter III. The classic treatment of random walks and the reflection principle.
- **Richard Stanley, *Catalan Numbers*** (Cambridge University Press). Over two hundred interpretations, with exercises.
- **Marc Renault, "Lost (and found) in translation: André's actual method and its application to the generalized ballot problem"**, *American Mathematical Monthly* 115 (2008). The history of the reflection principle.



# 9.11 Collecting the Set

## The question

The makers of Cruncho (“The Cereal Which Cares”) are giving away a series of cards depicting  $n$  great mathematicians. Each packet of Cruncho contains one picture chosen at random. Show that when I have collected  $r$  different cards the expected number of packets I must open to find a new card is  $n/(n-r)$  where  $0 \leq r \leq n-1$ .

Show by means of a diagram, or otherwise, that

$$\frac{1}{r+1} \leq \int_r^{r+1} \frac{1}{x} dx \leq \frac{1}{r}$$

and deduce that

$$\sum_{r=2}^n \frac{1}{r} \leq \ln n \leq \sum_{r=1}^{n-1} \frac{1}{r}$$

for all  $n \geq 2$ .

My children will give me no peace until we have the complete set of cards, but I am the only person in our household prepared to eat Cruncho and my spouse will only buy the stuff if I eat it. If  $n$  is large, roughly how many packets must I expect to consume before we have the set?

## The hook

A football sticker album has 600 stickers to collect, sold one at a time at random. How many stickers should you expect to buy to fill the album? The first few hundred go quickly, since almost every sticker is new. But as the album fills, new stickers get rarer, and the very last one takes on average 600 purchases on its own. The total comes to about 4200, seven times the size of the album.

This is the *coupon collector's problem*. Its answer combines two ideas: the waiting time for a success, and the harmonic series  $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ , which grows like  $\ln n$ . You met the harmonic series diverging in Section 2.3; here its slow growth is exactly what you pay for completing a collection.

## Before you start



**Ask yourself** If each packet independently gives a new card with probability  $p$ , what is the expected number of packets you open up to and including the first new card?

The number of packets  $X$  has a *geometric distribution*:  $P(X = k) = (1 - p)^{k-1}p$ . Its expectation is  $\frac{1}{p}$ . (One quick way to see this:  $E(X) = 1 + (1 - p)E(X)$ , since you open one packet, and with probability  $1 - p$  you are back where you started.) If a new card turns up a third of the time, you expect to wait three packets.

## Working it through

### STAGE 1 Waiting for a new card

With  $r$  different cards already, a packet gives a new card with probability  $\frac{n-r}{n}$ , independently of previous packets. So the expected number of packets until the next new card is

$$\frac{1}{(n-r)/n} = \frac{n}{n-r}. \quad \blacksquare$$

### STAGE 2 Comparing a sum with an integral



**Ask yourself** On the interval  $r \leq x \leq r+1$ , between which two values does  $\frac{1}{x}$  lie? What does that say about the area under the curve?

Since  $\frac{1}{x}$  is decreasing,  $\frac{1}{r+1} \leq \frac{1}{x} \leq \frac{1}{r}$  on  $[r, r+1]$ . Integrating over an interval of length 1 gives

$$\frac{1}{r+1} \leq \int_r^{r+1} \frac{dx}{x} \leq \frac{1}{r}. \quad \blacksquare$$

In pictures (Figure 9.11.1): the area under the curve is squeezed between a rectangle of height  $\frac{1}{r+1}$  and one of height  $\frac{1}{r}$ . Adding these from  $r = 1$  to  $r = n - 1$ , the integrals join up to  $\int_1^n \frac{dx}{x} = \ln n$ , and so

$$\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \leq \ln n \leq 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}. \quad \blacksquare$$

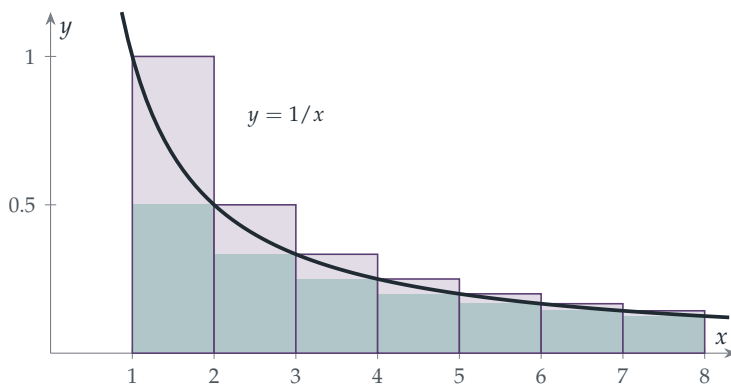


**Figure 9.11.1** On each interval  $[r, r + 1]$  the area under  $y = 1/x$  lies between the teal rectangle (height  $\frac{1}{r+1}$ ) and the larger plum one (height  $\frac{1}{r}$ ). Adding up,  $\ln n$  lies between  $\frac{1}{2} + \cdots + \frac{1}{n}$  and  $1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$ . The gaps between the curve and the plum rectangles add up to a finite total: Euler's constant  $\gamma \approx 0.577$ .

### STAGE 3 Completing the set



**Ask yourself** The total number of packets is the sum of the waiting times for the first card, the second new card, and so on. What is its expectation?

Let  $X_r$  be the number of packets needed to go from  $r$  different cards to  $r + 1$ . The total is  $X_0 + X_1 + \cdots + X_{n-1}$ , and since expectation adds,

$$E(\text{total}) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n \left( \frac{1}{n} + \frac{1}{n-1} + \cdots + 1 \right) = nH_n,$$

where  $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$  is the  $n$ th harmonic number. From Stage 2,  $\ln n + \frac{1}{n} \leq H_n \leq \ln n + 1$ , so  $H_n$  is about  $\ln n$  when  $n$  is large, and I must expect to eat roughly

$n \ln n$  packets. ■

### Looking back

- **Break a long wait into stages.** The total time split into  $n$  waiting times, each of a simple geometric kind.
- **Expectation is additive.** We never needed the distribution of the total, only the expectations of the pieces.
- **Compare sums with integrals.** The diagram turned a sum we cannot evaluate into a logarithm, with error at most 1.

### What's really going on

### How good is the estimate?

The comparison can be sharpened. The shaded gaps between the curve and the upper rectangles in Figure 9.11.1 have a finite total area, so  $H_n - \ln n$  tends to a limit,

$$\gamma = \lim_{n \rightarrow \infty} (H_n - \ln n) = 0.5772156649 \dots,$$

*Euler's constant.* (Nobody knows whether  $\gamma$  is rational.) So the expected number of packets is  $n \ln n + \gamma n + \frac{1}{2} + \dots$ . For the 600-sticker album,  $600 \times (\ln 600 + 0.577) \approx 4185$ , which is the figure quoted in the hook.

Where does the time go? The stages are very unequal. Collecting the first half of the set takes only about  $n \ln 2 \approx 0.69n$  packets, while the last card alone takes  $n$  on average. Most of the effort goes into the last few cards, which is why collectors swap duplicates.

### Concentration

The expected value is only the start: how much does the total vary? The variance of the total is  $\sum nr / (n - r)^2$ , which is about  $\frac{\pi^2}{6} n^2$  (the sum  $\sum 1/k^2 = \pi^2/6$  from Section 2.4 appears). So the standard deviation is about  $1.28n$ , much smaller than the mean  $n \ln n$  when  $n$  is large: the total is *concentrated* around its mean. Paul Erdős and Alfréd Rényi showed in 1961 that, more precisely, the total minus  $n \ln n$ , divided by  $n$ , has a limiting distribution, the *Gumbel distribution*, which also describes the largest of many random quantities (Section 12.25).

### Coupons everywhere

The problem goes back to Abraham de Moivre (1712) and Laplace. It now appears in many disguises. How many random tests are needed to check every component of a system? How long until a random walk on a network has visited every node? How many times must a computer hash table be probed? In each case the answer involves  $n \ln n$ . In biology, the same calculation estimates how many sequencing “reads” are needed to cover a whole genome.

## The thinking skill

### Linearity of expectation

*The expectation of a sum is the sum of the expectations, whether or not the pieces are independent. Split a complicated random quantity into simple pieces.*

This is one of the most useful facts in probability. Here it reduced the coupon collector's problem to  $n$  geometric waits; in Section 9.8 it gave the expected number of matches without any calculation. At university, the same idea powers the analysis of algorithms and the probabilistic method in combinatorics.

## Try this next

---

- 1 **Half a set.** Show that the expected number of packets to collect  $\frac{n}{2}$  different cards (for even  $n$ ) is  $n(H_n - H_{n/2})$ , and that this is about  $n \ln 2$ .
- 2 **Distinct cards.** After opening  $k$  packets, show that the expected number of different cards you have is  $n(1 - (1 - \frac{1}{n})^k)$ . (*Hint:* for each card, find the probability that you have it, and add.) Roughly how many packets give you 95% of the set?
- 3 **Two collectors.** Two children collect the same set, sharing packets so that a duplicate for one may be new for the other. Show that the expected number of packets to complete both sets is about  $n(\ln n + \ln \ln n)$ . (This is harder; first convince yourself that it is much less than  $2n \ln n$ .)
- 4 **A sharper bound.** Use the trapezium rule picture instead of rectangles to show that  $H_n - \ln n$  is decreasing, and that  $H_n - \ln n - \frac{1}{2n}$  is increasing. Deduce  $0.5 < \gamma < 0.7$  from small values of  $n$ .

## Further reading

---

- **Geoffrey Grimmett & David Stirzaker, *One Thousand Exercises in Probability*** (Oxford University Press). Many variations on the coupon collector.
- **Julian Havil, *Gamma: Exploring Euler's Constant*** (Princeton University Press). Everything about  $\gamma$  and the harmonic series, at an accessible level.
- **Michael Mitzenmacher & Eli Upfal, *Probability and Computing*** (Cambridge University Press). The coupon collector and its uses in computer science.



# 9.12 Random Subsets

## The question

The set  $S$  is the set of all integers from 1 to  $n$ . The set  $T$  is the set of all distinct subsets of  $S$ , including the empty set  $\emptyset$  and  $S$  itself. Show that  $T$  contains exactly  $2^n$  sets.

The sets  $A_1, A_2, \dots, A_m$ , which are not necessarily distinct, are chosen randomly and independently from  $T$ , and for each  $k$  ( $1 \leq k \leq m$ ), the set  $A_k$  is equally likely to be any of the sets in  $T$ .

- (i) Write down the value of  $P(1 \in A_1)$ .
- (ii) By considering each integer separately, show that  $P(A_1 \cap A_2 = \emptyset) = \left(\frac{3}{4}\right)^n$ .  
Find  $P(A_1 \cap A_2 \cap A_3 = \emptyset)$  and  $P(A_1 \cap A_2 \cap \dots \cap A_m = \emptyset)$ .
- (iii) Find  $P(A_1 \subseteq A_2)$ ,  $P(A_1 \subseteq A_2 \subseteq A_3)$  and  $P(A_1 \subseteq A_2 \subseteq \dots \subseteq A_m)$ .

## The hook

Choose two subsets of  $\{1, 2, \dots, 100\}$  completely at random. How likely is it that they have no element in common? Your intuition may say “quite likely: most pairs of sets are different”. In fact the probability is  $\left(\frac{3}{4}\right)^{100}$ , about  $3 \times 10^{-13}$ . Random sets are big (about half of everything) and so they almost always overlap.

The question is short, and its one idea is powerful: do not think about the sets as a whole; think about one element at a time. A random subset is just  $n$  independent coin tosses, one for each element, and every question about random subsets becomes a question about those tosses.

## Before you start



**Ask yourself** How can you describe a subset of  $\{1, \dots, n\}$  by a sequence of  $n$  answers to yes/no questions?

For each element  $i$ , answer the question “is  $i$  in the set?”. The subset determines the answers, and the answers determine the subset. There are 2 choices for each element, so  $2^n$  subsets in all. Choosing a subset uniformly at random is the same as answering each question by tossing a fair coin, independently.

## Working it through

### STAGE 1 One element

By the coin-tossing description,  $P(1 \in A_1) = \frac{1}{2}$ . Indeed, exactly half of the  $2^n$  subsets contain 1: pair each subset that does not contain 1 with the one obtained by adding 1.

### STAGE 2 Disjoint sets



**Ask yourself** For  $A_1 \cap A_2 = \emptyset$ , what must be true of each element  $i$ ? What is the probability of that, for one element? Why can you multiply over the elements?

$A_1$  and  $A_2$  are disjoint exactly when no element is in both. For a single element  $i$ , there are four equally likely possibilities: in neither set, in  $A_1$  only, in  $A_2$  only, in both (Figure 9.12.1, left). Only the last is forbidden, so the probability that  $i$  is not in both is  $\frac{3}{4}$ . Different elements are decided by different, independent coin tosses, so

$$P(A_1 \cap A_2 = \emptyset) = \left(\frac{3}{4}\right)^n. \quad \blacksquare$$

For three sets, element  $i$  is in all three with probability  $\frac{1}{8}$ , so  $P(A_1 \cap A_2 \cap A_3 = \emptyset) = \left(\frac{7}{8}\right)^n$ . In general,

$$P(A_1 \cap \cdots \cap A_m = \emptyset) = \left(1 - \frac{1}{2^m}\right)^n. \quad \blacksquare$$

### STAGE 3 Subsets



**Ask yourself** For  $A_1 \subseteq A_2$ , which of the four possibilities for an element is forbidden?

$A_1 \subseteq A_2$  fails exactly when some element is in  $A_1$  but not in  $A_2$ . Again one pattern out of four is forbidden, so

$$P(A_1 \subseteq A_2) = \left(\frac{3}{4}\right)^n,$$

the same as for disjointness. That is not a coincidence:  $A_1 \subseteq A_2$  exactly when  $A_1$  is disjoint from the complement of  $A_2$ , and the complement of a random set is another random set.



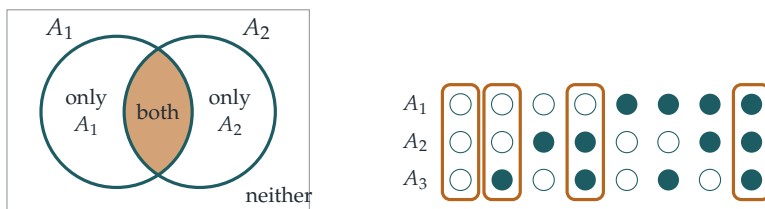
**Ask yourself** For  $A_1 \subseteq A_2 \subseteq A_3$ , which of the 8 patterns of membership (in or out of each set) are allowed for an element?

Going along the chain  $A_1, A_2, A_3$ , once an element is in, it must stay in. So the allowed patterns are “out, out, out”, “out, out, in”, “out, in, in” and “in, in, in”: 4 out of 8 (Figure 9.12.1, right). Hence  $P(A_1 \subseteq A_2 \subseteq A_3) = \left(\frac{4}{8}\right)^n = \left(\frac{1}{2}\right)^n$ .

For a chain of  $m$  sets, an allowed pattern is determined by the first set the element

enters (one of  $A_1, \dots, A_m$ , or never), so  $m + 1$  of the  $2^m$  patterns are allowed:

$$P(A_1 \subseteq A_2 \subseteq \dots \subseteq A_m) = \left(\frac{m+1}{2^m}\right)^n. \quad \blacksquare$$



each element: 4 equally likely places      allowed for  $A_1 \subseteq A_2 \subseteq A_3$ : 4 of 8

**Figure 9.12.1** Left: an element of  $S$  can be in one of four places relative to  $A_1$  and  $A_2$ , each with probability  $\frac{1}{4}$ . Disjointness forbids “both” (amber);  $A_1 \subseteq A_2$  forbids “only  $A_1$ ”. Right: the eight patterns of membership of  $A_1, A_2, A_3$  (filled circle means “in”). The chain  $A_1 \subseteq A_2 \subseteq A_3$  allows only the four boxed patterns, in which an element, once in, stays in.

## Looking back

- **Decompose by elements.** Each set-level condition (disjoint, subset, chain) was a condition on each element separately.
- **Independence across elements.** Because the elements are decided by independent coin tosses, the probabilities multiply.
- **Count patterns.** Each answer was (allowed patterns  $\div$  all patterns) $^n$ .

## What’s really going on

### Counting without probability

Multiplying by the number of choices turns each probability into a count. The number of ordered pairs  $(A_1, A_2)$  with  $A_1 \subseteq A_2$  is  $4^n \times \left(\frac{3}{4}\right)^n = 3^n$ . You can also count these pairs directly: choose the size  $k$  of  $A_2$ , then  $A_2$  itself in  $\binom{n}{k}$  ways, then  $A_1$  as any of its  $2^k$  subsets. So

$$\sum_{k=0}^n \binom{n}{k} 2^k = 3^n,$$

which is the binomial theorem for  $(1 + 2)^n$ . Counting the same thing in two ways, once by elements and once by sizes, proves an identity: this is the method of *double counting*, which you met with the Platonic solids in Section 7.9. Similarly, the number of chains  $A_1 \subseteq \dots \subseteq A_m$  is  $(m + 1)^n$ .

## The lattice of subsets

The subsets of  $\{1, \dots, n\}$ , ordered by inclusion, form a structure called the *Boolean lattice*, which can be pictured as an  $n$ -dimensional cube: each subset is a corner, labelled by its sequence of yes/no answers, and moving along an edge adds or removes one element. Chains are paths going upwards. Deep results describe how large families of subsets can be, given conditions on them. *Sperner's theorem* (1928) says that the largest family in which no set contains another has  $\binom{n}{\lfloor n/2 \rfloor}$  sets; the *Erdős–Ko–Rado theorem* (1961) says how many  $k$ -element sets can be chosen so that every two overlap. Both are standard in a first course on combinatorics, and the coin-tossing view of subsets gives some of the slickest proofs.

### Boole's algebra

George Boole's *The Laws of Thought* (1854) treated logical statements as sets and “and”, “or” and “not” as intersection, union and complement. Encoding a set as a string of yes/no answers is exactly how a computer represents it: a subset of  $\{1, \dots, 64\}$  is a single 64-bit number, and intersection is the “bitwise AND” instruction. The method of this question, one element (one bit) at a time, is how such operations are analysed.

## The thinking skill

### Think one element at a time

*When objects are built from independent choices for each element, a condition that applies to every element separately has probability equal to a product.*

The method turned a question about  $2^n$  subsets into a question about 4 or 8 patterns. At university it reappears in combinatorics (generating functions for sets and multisets), in probability (independent coordinates) and in computer science (bitwise reasoning).

## Try this next

- Covering.** Find  $P(A_1 \cup A_2 = S)$  and  $P(A_1 \cup A_2 \cup \dots \cup A_m = S)$ .
- Same size.** Show that  $P(|A_1| = |A_2|) = \binom{2n}{n}/4^n$ . (*Hint:*  $|A_1| = |A_2|$  exactly when  $|A_1| + |S \setminus A_2| = n$ .) Show that this is about  $1/\sqrt{\pi n}$  for large  $n$ .
- Partitions into three.** Show that the number of ordered pairs of disjoint subsets  $(A, B)$  of  $S$  is  $3^n$ , and find the number of unordered pairs  $\{A, B\}$  of disjoint subsets.
- Sperner for small  $n$ .** For  $n = 3$  and  $n = 4$ , find the largest family of subsets of  $\{1, \dots, n\}$  in which no set contains another.

## Further reading

---

- ▶ **Miklós Bóna**, *A Walk Through Combinatorics* (World Scientific), the chapters on subsets and partially ordered sets.
- ▶ **Béla Bollobás**, *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability* (Cambridge University Press). A short, elegant book on families of subsets, from Sperner's theorem onwards.
- ▶ **Peter Cameron**, *Combinatorics: Topics, Techniques, Algorithms* (Cambridge University Press). A first course with a strong emphasis on counting in two ways.



# 9.13 Loaded Dice

## The question

- (i) Prove that if  $x > 0$  then  $x + x^{-1} \geq 2$ .

I have a pair of six-faced dice, each with faces numbered from 1 to 6. The probability of throwing  $i$  with the first die is  $q_i$  and the probability of throwing  $j$  with the second die is  $r_j$  ( $1 \leq i, j \leq 6$ ). The two dice are thrown independently and the sum noted. By considering the probabilities of throwing 2, 12 and 7, show the sums  $2, 3, \dots, 12$  are not equally likely.

- (ii) The first die described above is thrown twice and the two numbers on the die noted. Is it possible to find values of  $q_j$  so that the probability that the numbers are the same is less than  $1/36$ ?

## The hook

With two fair dice, a total of 7 is six times as likely as a total of 2. Board-game designers sometimes wish it were otherwise: could you load the dice (weight them so that some faces come up more often) so that every total from 2 to 12 has the same chance,  $\frac{1}{11}$ ? You can choose the twelve probabilities freely, so there seems to be plenty of room.

The answer is no, and the proof is a gem: it uses only three of the eleven totals and the inequality  $x + \frac{1}{x} \geq 2$ . It is a model of how to prove that something is *impossible*, which is often harder, and more interesting, than showing that something can be done.

## Before you start



**Ask yourself** How would you prove  $x + \frac{1}{x} \geq 2$  for  $x > 0$ ? When is there equality?

Multiply by  $x > 0$ : the inequality becomes  $x^2 + 1 \geq 2x$ , that is  $(x - 1)^2 \geq 0$ , which is true. Equality holds only when  $x = 1$ . (Alternatively,  $x + \frac{1}{x} - 2 = (\sqrt{x} - \frac{1}{\sqrt{x}})^2$ .) This is the AM–GM inequality in its simplest form: the average of  $x$  and  $\frac{1}{x}$  is at least their geometric mean, 1.

## Working it through

### STAGE 1 The extreme totals



**Ask yourself** In how many ways can you throw a total of 2? A total of 12? Write their probabilities in terms of the  $q_i$  and  $r_j$ .

A total of 2 needs 1 on both dice, and 12 needs 6 on both, so  $P(2) = q_1r_1$  and  $P(12) = q_6r_6$ .

Suppose, for contradiction, that all eleven totals are equally likely, each with probability  $\frac{1}{11}$ . Then

$$q_1r_1 = \frac{1}{11}, \quad q_6r_6 = \frac{1}{11}.$$

In particular, all four of  $q_1, r_1, q_6, r_6$  are positive.

### STAGE 2 The middle total



**Ask yourself** A total of 7 can be made in six ways. Which two of them involve only the faces 1 and 6? Write their probabilities using the equations above.

The total 7 includes the outcomes (1, 6) and (6, 1), so

$$P(7) \geq q_1r_6 + q_6r_1.$$

Now use the equations of Stage 1 to bring in a ratio. Since  $q_1 = \frac{1}{11r_1}$  and  $q_6 = \frac{1}{11r_6}$ ,

$$q_1r_6 + q_6r_1 = \frac{1}{11} \left( \frac{r_6}{r_1} + \frac{r_1}{r_6} \right) \geq \frac{1}{11} \times 2 = \frac{2}{11},$$

using  $x + \frac{1}{x} \geq 2$  with  $x = r_6/r_1 > 0$ . So  $P(7) \geq \frac{2}{11}$ , contradicting  $P(7) = \frac{1}{11}$ . The totals cannot all be equally likely. ■



**Pause and try** Look at what made the proof work. Making 2 and 12 as likely as 7 forces the 1s and 6s to be fairly likely, and then the “mixed” outcomes (1, 6) and (6, 1) alone make 7 too likely. Figure 9.13.1 shows the cells involved.

|            |   | first die |   |   |    |    |    |
|------------|---|-----------|---|---|----|----|----|
|            |   | 1         | 2 | 3 | 4  | 5  | 6  |
| second die | 1 | 2         | 3 | 4 | 5  | 6  | 7  |
|            | 2 | 3         | 4 | 5 | 6  | 7  | 8  |
|            | 3 | 4         | 5 | 6 | 7  | 8  | 9  |
|            | 4 | 5         | 6 | 7 | 8  | 9  | 10 |
|            | 5 | 6         | 7 | 8 | 9  | 10 | 11 |
|            | 6 | 7         | 8 | 9 | 10 | 11 | 12 |

■ totals 2 and 12:  $q_1r_1, q_6r_6$   
■ (1,6) and (6,1):  $q_1r_6, q_6r_1$   
■ the rest of total 7

**Figure 9.13.1** The 36 outcomes of two dice, labelled by their totals. The proof uses only the four corner cells. The product of the two amber probabilities equals the product of the two dark teal ones,  $q_1r_1q_6r_6 = q_1r_6 \cdot q_6r_1$ , and by AM–GM the sum of the teal pair is at least twice the geometric mean of the amber pair.

### STAGE 3 Two throws of one die



**Ask yourself** If one die is thrown twice, what is the probability that the two numbers are the same, in terms of the  $q_j$ ? What is it for a fair die?

The two throws agree with probability  $q_1^2 + q_2^2 + \cdots + q_6^2$ . For a fair die this is  $6 \times \frac{1}{36} = \frac{1}{6}$ , which is already much bigger than  $\frac{1}{36}$ .



**Ask yourself** Can loading make  $\sum q_j^2$  smaller than  $\frac{1}{6}$ ? You know that  $\sum q_j = 1$ .

Write  $q_j = \frac{1}{6} + e_j$ , where the deviations  $e_j$  add to zero. Then

$$\sum q_j^2 = \sum \left( \frac{1}{36} + \frac{e_j}{3} + e_j^2 \right) = \frac{1}{6} + \frac{1}{3} \sum e_j + \sum e_j^2 = \frac{1}{6} + \sum e_j^2 \geq \frac{1}{6}.$$

So the probability of a repeat is always at least  $\frac{1}{6}$ , with equality only for a fair die. It certainly cannot be less than  $\frac{1}{36}$ . ■

### Looking back

- **Proof by contradiction.** We assumed all totals equally likely and derived an impossibility.
- **Use only what you need.** Of the eleven equations, three were enough. Choosing the right ones was the key step.
- **Inequalities prove impossibility.** AM–GM in part (i), and “a sum of squares is non-negative” in part (ii), ruled out whole families of possibilities at once.

## What's really going on

---

### A proof with polynomials

Here is a completely different proof, using the generating functions of Section 9.6. Let  $Q(x) = q_1x + \cdots + q_6x^6$  and  $R(x) = r_1x + \cdots + r_6x^6$ . If all totals had probability  $\frac{1}{11}$ , then

$$Q(x)R(x) = \frac{x^2 + x^3 + \cdots + x^{12}}{11} = \frac{x^2}{11}(1 + x + \cdots + x^{10}).$$

Now  $Q(x)/x$  is a polynomial of degree 5 with real coefficients, and every real polynomial of odd degree has a real root (its graph goes from  $-\infty$  to  $+\infty$ , or the other way). So the left side has a real non-zero root. But  $1 + x + \cdots + x^{10} = \frac{x^{11}-1}{x-1}$ , whose roots are the 11th roots of unity other than 1, none of which is real. Contradiction. The same argument shows that no loading of two  $n$ -sided dice can make the totals equally likely whenever  $n-1$  is odd, that is, for  $n$  even. (For odd  $n$  the polynomial argument needs more work, but the AM–GM proof above works for every  $n \geq 2$ .)

### Collision probability

The quantity  $\sum q_j^2$  in part (ii), the chance that two independent draws agree, is called the *collision probability* of a distribution. Part (ii) shows that it is smallest for the uniform distribution. It measures how concentrated a distribution is: in ecology,  $\sum q_j^2$  is Simpson's index of diversity, where  $q_j$  is the proportion of species  $j$  in a habitat; in economics, it is the Herfindahl index of market concentration; in cryptography, it measures how guessable a password distribution is. And it explains the remark at the end of Section 9.9: unequal birthday probabilities can only make shared birthdays more likely.

### Impossibility proofs

Some of the most famous results in mathematics say that something cannot be done: no fraction squares to 2; no ruler-and-compass construction trisects every angle; no formula in radicals solves every quintic; no algorithm decides every mathematical statement. Each changed its subject, because to prove that *no* method works, you must understand the whole space of methods. The loaded dice problem is a miniature: to rule out every loading, we needed an invariant property (the AM–GM inequality) that every loading must obey.

---

## The thinking skill

### Prove impossibility by contradiction and inequality

*To show that no choice of parameters works, assume one does, pick out the few conditions that clash, and use an inequality to show they cannot all hold.*

At university, contradiction combined with inequalities is the standard way to prove that objects do not exist, in analysis, in combinatorics and in optimisation, where inequalities like AM–GM and Cauchy–Schwarz set limits on what can be achieved.

## Try this next

- 1 Cauchy–Schwarz.** Use the Cauchy–Schwarz inequality,  $(\sum a_j b_j)^2 \leq \sum a_j^2 \sum b_j^2$ , with  $a_j = q_j$  and  $b_j = 1$ , to give a one-line proof of part (ii).
- 2 Two different dice.** If two *different* loaded dice are each thrown once, can the probability that they show the same number be less than  $\frac{1}{6}$ ? Can it be 0?
- 3 Uniform totals with other faces.** Show that two dice *can* give equally likely totals if their faces are numbered differently: try one die numbered 0, 1, 2, 3, 4, 5 and the other 0, 6, 12, 18, 24, 30. What totals arise? Why does this not contradict the question?
- 4 Three dice.** Can three loaded six-sided dice, with faces 1 to 6, make all totals from 3 to 18 equally likely? (*Hint:* use the totals 3, 8, 13 and 18, and AM–GM for six numbers. Why does the polynomial proof fail this time?)

## Further reading

- **J. Michael Steele, *The Cauchy–Schwarz Master Class*** (Cambridge University Press). A beautiful introduction to inequalities, starting from AM–GM.
- **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*** (Oxford University Press). Generating functions and their use in proofs like the one above.
- **Ian Stewart, *Galois Theory*** (CRC Press), the opening chapters. The most famous impossibility proof of all, for the quintic.



# What We Have Learned

---

We began by saying that counting is the oldest mathematics there is, and that doing it cleverly is still one of the hardest. We asked three questions: how do you count without listing, when must something happen, and how do counting and probability help each other? Here is what the chapter's thirteen questions showed.

## 1. How do you count without listing?

The chapter built up a toolkit.

- ▶ **Recurrences.** Adding one line to the plane (Section 9.1) or one post to a fence (Section 9.3) changed the count in a simple way, giving a recurrence that could be solved. Recording a little extra information, the colour of the last post, made the recurrences close up.
- ▶ **Generating functions.** Packaging a sequence as a power series turned recurrences into algebra (Section 9.4), explained why  $1/89 = 0.01123\dots$ , and turned counting combinations into multiplying polynomials: coins in Section 9.5, dice in Section 9.6, and balanced power sums in Section 9.7, where a repeated root at  $x = 1$  did all the work.
- ▶ **Bijections.** The reflection principle (Section 9.10) counted bad queues by matching them with unrestricted paths, and complements of random sets turned subsets into disjoint pairs (Section 9.12).
- ▶ **Inclusion–exclusion.** Overcounting and correcting found the number of derangements (Section 9.8).
- ▶ **Double counting.** Counting regions with Euler's formula (Section 9.1) and pairs of subsets by sizes (Section 9.12) gave the same answers by different routes, and proved identities on the way.

## 2. When must something happen?

Several sections showed that size alone forces structure. Among six people there must be three mutual friends or three mutual strangers (Section 9.2), by the pigeonhole principle, though five people are not enough. Twenty-three people make a shared birthday more likely than not (Section 9.9), because the number of pairs grows like the square of the number of people. Random subsets of a large set almost always overlap (Section 9.12). And some things *cannot* happen, however you try: no loading of two dice makes every total equally likely (Section 9.13), and no numbering of dice other than Sicherman's reproduces the ordinary totals (Section 9.6).

3. How do counting and probability help each other?

In the second half of the chapter, probabilities were found by counting: derangements by inclusion–exclusion, the birthday problem by a product, the ballot problem by reflection. But the traffic went the other way too. Linearity of expectation gave the coupon collector’s  $nH_n$  (Section 9.11) and the expected number of matches in a shuffle without any counting at all. Thinking of a random subset as  $n$  coin tosses (Section 9.12) turned counts like  $3^n$  and  $(m + 1)^n$  into one-line calculations. And Erdős’s probabilistic method (Section 9.2) proved that good colourings exist by showing that a random colouring works with positive probability. Two constants appeared from nowhere:  $e$ , in the limiting chance of a derangement, and Euler’s  $\gamma$ , in the cost of completing a collection.

4. A toolkit for counting

| Method                                   | Where you used it                      |
|------------------------------------------|----------------------------------------|
| What changes when you add one            | Lines (9.1); fence posts (9.3)         |
| Pigeonhole principle                     | Ramsey (9.2)                           |
| Coupled recurrences                      | Fence posts (9.3)                      |
| Generating functions                     | 9.4, 9.5, 9.6, 9.7, 9.13               |
| Case analysis from the extremes          | Strange dice (9.6); loaded dice (9.13) |
| Inclusion–exclusion                      | Derangements (9.8)                     |
| Complement and exponential approximation | Birthdays (9.9)                        |
| Reflection principle                     | Ballot problem (9.10)                  |
| Linearity of expectation                 | Coupons (9.11); matches (9.8)          |
| One element at a time                    | Random subsets (9.12)                  |
| Contradiction with AM–GM                 | Loaded dice (9.13)                     |

Where next

- The Poisson limit of derangements returns in Section 12.11, and generating functions of random variables drive the study of family trees in Section 12.15.
- Coupled recurrences become Markov chains in Section 12.16.
- The Gumbel distribution of the coupon collector describes the largest of many random quantities in Section 12.25.

If you want to go further before university

- **Miklós Bóna, *A Walk Through Combinatorics***: a friendly first course covering almost every idea in this chapter.
- **Herbert Wilf, *generatingfunctionology***: the classic on generating functions; free online.

- ▶ **Ronald Graham, Donald Knuth & Oren Patashnik, *Concrete Mathematics*:** sums, recurrences and generating functions, with wit and many exercises.
- ▶ **Geoffrey Grimmett & Dominic Welsh, *Probability: An Introduction*:** the counting side of first-year probability.

Whitehead called combinatorics “the slums of topology”. This chapter has tried to show why nobody would say that now. Behind every clever count there was an idea that reaches far beyond it: a generating function, a bijection, a recurrence, an inequality. Those ideas connect counting with algebra, analysis and probability, and learning to look for them, rather than for one more trick, is what turns counting into mathematics.