The Probabilistic Method

David Arthur*
darthur@gmail.com

Sometimes you will run across Olympiad problems where the main challenge is to show that there exists an object with some specific property. If you're lucky, you may just be able to write down what the object should be, and then observe it has the desired property. Often though, that is out of the question. There might be too many cases to consider, or the object you are looking for might be too complicated. So what do you do when that happens? In this talk, I am going to focus on one particular option:

The Probabilistic Method:

1. Choose an object at random.

   The basic method: Choose an object *uniformly at random*. This means that each object is equally likely to be picked.[1]

   Variations: You do not *have* to choose an object uniformly at random. You just need to describe a way of picking an object that has some randomness involved. As you will see, there are sometimes other effective ways of doing this.

2. Prove that you will choose an object with the desired property with probability $p > 0$.

   Method #1: Often you will be looking for an object where some numeric quantity X has value at least y (or value at most y). If you can show that, *on average*, your random object will have value $X = y$, it follows immediately that in some case $X \geq y$, and in some case $X \leq y$.

   Method #2: You can also try to calculate the probability that your random object will *not* have the desired property. If you can prove this probability is less than 1 with inequality techniques, you are done.

*Based partially on notes by Po-Shen Loh.

[1]Usually there are only finitely many objects to choose from, in which case this makes perfect sense. If there are an infinite number of objects, things become more complicated. In particular, there is no such thing as choosing something uniformly at random from the set of all real numbers or from the set of all integers. You can choose a real number uniformly at random from a *finite* interval, or a point uniformly at random from a figure with *finite* area. The probability that a point lies in some region is then proportional to the length (or area) of that region.

Of course, anything that can be done with the probabilistic method can also be done with basic combinatorics. Method #1 combines counting with the Pigeonhole Principle, and method #2 combines counting with inequalities. Working in terms of probabilities often makes things simpler though, and it definitely helps guide your thinking. The techniques here are widely used in research mathematics and computer science, which should give you some indication that they are an effective and elegant way of looking at the world!

Anyway, let's start with a basic example.

Example 1. *At the IMO, there are n students, and m pairs of these students are enemies. Prove that it is possible to divide the students into k rooms so that there are at most $\frac{m}{k}$ pairs of enemies that are placed in the same room as each other.*

Solution. Independently place each student in a random room. Now consider a fixed pair of enemies. The probability that they will be placed in the same room is exactly $\frac{1}{k}$. Therefore, on *average*, the number of pairs of enemies placed in the same room is exactly $\frac{m}{k}$. (Do you see why?) In particular, there exists some configuration where at most $\frac{m}{k}$ enemies are placed in the same room. □

This is a pretty easy problem, so you could also solve it by the extremal principle or by a greedy assignment. If you tried to directly translate the probabilistic argument into a counting argument though, things would get messy.

1 Calculating probabilities

When applying the probabilistic method, it is very important to be comfortable with calculating probabilities! Here are a few facts that are particularly helpful. The first two facts you are probably already familiar with, but the third may be new:

- Let A and B be two random events. Then,

$$\text{Prob}[A \textit{ or } B] = \text{Prob}[A] + \text{Prob}[B] - \text{Prob}[A \textit{ and } B].$$

  In particular, $\text{Prob}[A \textit{ or } B] \leq \text{Prob}[A] + \text{Prob}[B]$. This is called the *union bound*, and equality holds if and only if A and B are mutually exclusive.

- If A and B are two random variables, we say A and B are *independent* if knowing A gives no information about B. For example, if I roll two dice, the rolls are independent from each other, but neither roll is independent from the sum. Then, A and B being independent is equivalent to the following:

$$\text{Prob}[A = a \textit{ and } B = b] = \text{Prob}[A = a] \cdot \text{Prob}[B = b] \qquad \text{for all } a, b.$$

- For any probability $p \in [0, 1]$ and any positive integer n, we have $(1 - p)^n \leq e^{-np}$. This means that if $A_1, A_2, \ldots, A_n$ are random events that each happen independently with probability p, then at least one will happen with probability at least $1 - e^{-np}$.

Example 2. *At the IMO, there are n people, some of whom are students and some of whom are guides. Each person brought $k > \log_2 n$ different colored shirts. To avoid confusion, the IMO wants to ensure that no guide is wearing the same colored shirt as a student. Prove that there is a choice of shirts which ensures this.*

Solution. For each shirt color, we choose randomly and independently whether it will be allowed for students or allowed for guides. We then need to show that with positive probability, every person has at least one shirt they can wear.

Towards that end, consider a single person. The person owns k shirt colors, and the probability that *all* of them are assigned to the other side is exactly $\left(\frac{1}{2}\right)^k$. By the union bound, it follows that the probability of *somebody* having no valid shirts is at most $n \cdot \left(\frac{1}{2}\right)^k < n \cdot \left(\frac{1}{2}\right)^{\log_2 n} = 1$. Therefore, the probability of everybody having at least one valid shirt is greater than 0, and we're done. □

2 Expected value

The example from the previous section used Method #2, but Method #1 comes up more often. The key to this is the idea of an *expected value*, which formalizes the "average" of some random variable.

Definition 1. *Let X be a random variable that takes values in some set S. Then, the* **expected value** *of S, denoted* $\mathrm{E}[X]$*, is*

$$\mathrm{E}[X] = \sum_{x \in S} x \cdot \mathrm{Prob}[X = x].$$

The following facts are useful:

- Let X be a random variable. Then, there is at least one case where $X \geq \mathrm{E}[X]$, and there is at least one case where $X \leq \mathrm{E}[X]$.
- If X and Y are random variables, then $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$, even if X and Y are not independent. This is called *linearity of expectation* and it is a lot more helpful than you might think at first glance!
- If X and Y are independent random variables, then $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$.

Exercise 1. *If a fair coin is flipped 100 times, what is the expected number of heads?*

Solution. Let X denote the total number of heads, and let X_i denote the number of heads on the i^{th} coin toss. Then, $X = \sum_{i=1}^{100} X_i$ and $\mathrm{E}[X_i] = \frac{1}{2}$ for all i, so linearity of expectation implies that $\mathrm{E}[X] = \sum_{i=1}^{100} \mathrm{E}[X_i] = 50$. □

Exercise 2. *A fair coin is flipped repeatedly until it turns up tails five times. What is the expected number of heads before that happens?*

Exercise 3. *At the Pseudo-IMO, there are n students other than you, each from a different country. Every day, you eat lunch with a random student, and they give you a miniature flag from their country. Let X denote the number of days before you have at least one flag from every country. What is* $\mathrm{E}[X]$*?*

With the background out of the way, let's try using these ideas to solve an Olympiad problem.

Example 3. (Russia 1996, #4) *In the Duma, there are 1600 delegates who have formed 16000 committees of 80 persons each. Prove that one can find two committees having at least four common members.*

Solution. Pick two committees independently and uniformly at random. Let X be the number of people in both committees, and let X_i be the $\{0,1\}$-random variable indicating whether the i^{th} person is in both chosen committees. By linearity of expectation,

$$\mathrm{E}[X] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \ldots + \mathrm{E}[X_{1600}].$$

The magic is that each $\mathrm{E}[X_i]$ is easy to calculate! Let n_i be the number of committees that the i^{th} person belongs to. Then, $\mathrm{E}[X_i] = \mathrm{Prob}[i^{\text{th}} \text{ person belongs to both committees}] = \binom{n_i}{2}/\binom{16000}{2}$. Note that $\sum_{i=1}^{1600} n_i$ is just the total size of all committees, which is $16000 \cdot 80$, so the average value of $\{n_i\}$ is precisely $\overline{n} = 800$. Since $\binom{x}{2}$ is convex, we can now apply Jensen's inequality:

$$\mathrm{E}[X] = \sum_{i=1}^{1600} \binom{n_i}{2}/\binom{16000}{2} \geq 1600 \cdot \binom{\overline{n}}{2}/\binom{16000}{2} = 1600 \cdot \frac{800 \cdot 799}{16000 \cdot 15999} > 3.995.$$

In particular, there exists some choice for which X is at least 3.995. Since X is an integer, it follows that $X \geq 4$ in this case, and we're done. □

Example 4. (Erdős, 1965) *A set S is called* sum-free *if there is no triple of (not necessarily distinct) elements $x, y, z \in S$ satisfying $x + y = z$. Prove that every set A of non-zero integers contains a subset $S \subseteq A$ of size $|S| > |A|/3$, which is sum-free.*

Solution. Let p be a prime number of the form $3k + 2$ such that p is greater than the maximum absolute value of any element in A.[2] Also let $C = \{k+1, k+2, \ldots, 2k+1\}$, and note that this set is sum-free modulo p. Pick r uniformly at random from $\{1, 2, \ldots, p-1\}$, and consider multiplying each element of A by r, modulo p.

Since p is large, every element in A is non-zero modulo p. Therefore, each element in A has probability exactly $\frac{|C|}{p-1} > \frac{1}{3}$ of mapping into C when multiplied by r. Linearity of expectation then implies that the expected number of elements mapping into C is $> \frac{|A|}{3}$. Let S be the set of these elements. We know there exists some choice of r for which $|S| \geq \mathrm{E}[|S|] > \frac{|A|}{3}$. It remains only to show that S is sum-free. Indeed, if $x, y, z \in S$, then $xr + yr \not\equiv zr \pmod{p}$ since $xr, yr, zr \in C$, and hence, $x + y \neq z$. □

[2]The existence of such a prime number p is implied by Dirichlet's prime number theorem, which states that for any relatively prime positive integers a and d, the arithmetic sequence $a, a+d, a+2d, \ldots$ contains infinitely many primes. This theorem is difficult to prove in general but there is an elementary proof for $a = 2, d = 3$. Can you find it?

3 Problems

The following problems can all be solved with the probabilistic method. In many cases, it simplifies and motivates things very nicely – in other cases, you may feel more comfortable with pure combinatorial thinking. Use whatever you are comfortable with! The last few problems are very difficult, and actually come from research mathematics, so beware!

1. *(Canadian MO 2009, #2)* Two circles of different radii are cut out of cardboard. Each circle is subdivided into 200 equal sectors. On each circle 100 sectors are painted white and the other sectors are painted black. The smaller circle is then placed on top of the larger circle, so that their centers coincide. Show that one can rotate the small circle so that the sectors on the two circles line up and at least 100 sectors on the small circle lie over sectors of the same color on the big circle.

2. *(MOP 2007)* In an $n \times n$ array, each of the numbers $1, 2, \ldots, n$ appears exactly n times. Show that there is a row or a column in the array with at least $\sqrt{n}$ distinct numbers.

3. *(Korean MO 2008, #6)* There is an $n \times n$ grid on a computer. Each of its n^2 squares displays an integer from 0 to k. For each of the n rows and each of the n columns, there is also a button that, if pressed, will increase every number in that row or column by 1. If a number ever reaches k, it immediately changes to 0. Initially, every square displayed 0, but then a number of buttons were pressed. Show that after at most kn more button presses, it is possible to change every number back to 0 again.

4. Let $v_1, v_2, \ldots, v_n$ be unit vectors in $\mathbb{R}^d$. Prove that it is possible to assign weights $\epsilon_i \in \{\pm 1\}$ such that the vector $\sum \epsilon_i v_i$ has Euclidean norm[3] less than or equal to $\sqrt{n}$.

5. *(IMO 1987, #3)* Let $x_1, x_2, \ldots, x_n$ be real numbers satisfying $x_1^2 + x_2^2 + \ldots + x_n^2 = 1$. Prove that for every integer $k \geq 2$, there are integers $a_1, a_2, \ldots, a_n$, not all zero, such that $|a_i| \leq k-1$ for all i, and
$$|a_1 x_1 + a_2 x_2 + \ldots + a_n x_n| \leq \frac{(k-1)\sqrt{n}}{k^n - 1}.$$

6. *(IMO Shortlist 1999, C4)* Let A be any set of n residues mod n^2. Show that there is a set B of n residues mod n^2 such that at least half of the residues mod n^2 can be written as $a + b$ with $a \in A$ and $b \in B$.

7. *(Iran TST 2008, #6)* Suppose 799 teams participate in a tournament in which every pair of teams plays against each other exactly once. Prove that there two disjoint groups A and B of 7 teams each such that every team from A defeated every team from B.

8. *(Austrian-Polish math competition 1997, #8)* Let n be a natural number and M a set with n elements. Find the largest integer k such that there exist k 3-element subsets of M, no two of which are disjoint.

[3]The Euclidean norm of a vector $v = (x_1, x_2, \ldots, x_d)$ is defined to be $\sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}$, and a unit vector is a vector with Euclidean norm equal to 1.

9. *(USAMO 1995, #5)* Suppose that in a certain society, each pair of persons can be classified as either amicable or hostile. We shall say that each member of an amicable pair is a friend of the other, and each member of a hostile pair is a foe of the other. Suppose that the society has n persons and q amicable pairs, and that for every set of three persons, at least one pair is hostile. Prove that there is at least one member of the society whose foes include $q(1-4q/n^2)$ or fewer amicable pairs.

10. *(Taiwan 1997, #9)* For $n \geq k \geq 3$, let $X = \{1, 2, \ldots, n\}$, and let $\mathcal{F}_k$ be a family of k-element subsets of X such that any two subsets in $\mathcal{F}_k$ have at most $k-2$ common elements. Show that there exists a subset M_k of X with $\lfloor \log_2 n \rfloor + 1$ elements containing no subset in $\mathcal{F}_k$.

11. *(Canadian Winter camp 2009)* We are given a collection T of circles of radius 1, which together cover an area S in the plane. Show that it is possible to choose a collection of non-overlapping circles from T which together cover an area that is greater than or equal to $\frac{\pi}{8\sqrt{3}} \cdot S$.

12. *(Romanian master in mathematics competition 2008, #4)* Prove that from among any $(n+1)^2$ points inside a square of side length positive integer n, one can pick three that form a triangle with area at most $\frac{1}{2}$.

13. *(Erdős-Ko-Rado theorem)* Let n, k be positive integers satisfying $n \geq 2k$, and let $\mathcal{C}$ be a collection of pairwise-intersecting k-element subsets of $\{1, 2, \ldots, n\}$. Prove that $|\mathcal{C}| \leq \binom{n-1}{k-1}$.

14. *(Sperner's other lemma)* Let $\mathcal{C}$ be a collection of subsets of $\{1, 2, \ldots, n\}$ such that no two distinct subsets $A, B \in \mathcal{C}$ satisfy $A \subseteq B$. Prove that $|\mathcal{C}| \leq \binom{n}{\lfloor n/2 \rfloor}$.

15. *(Bollobás 1965, and Summer camp 2008)* Let X be a finite set, and suppose $A_1, A_2, \ldots, A_m$, and $B_1, B_2, \ldots, B_m$ are subsets of X with $|A_i| = r$ and $|B_i| = s$ for all i. If $A_i \cap B_i = \emptyset$ for all i and $A_i \cap B_j \neq \emptyset$ for all $i \neq j$, prove that $m \leq \binom{r+s}{r}$.

16. *(Crossing lemma)* A graph with V vertices and E edges is drawn in the plane. Show that, as long as $E \geq 4V$, there will be at least $\frac{E^3}{64V^2}$ pairs of edges that cross.

17. *(Karger 1993)* In a connected, undirected graph, a collection of edges is called a **cut** if removing those edges will disconnect the graph. A cut is called a **min-cut** if there is no cut with fewer edges. Prove that the maximum number of min-cuts on a graph with n vertices is exactly $\binom{n}{2}$.

4 Selected hints

11. Tile the plane with circles and then add randomness. By the way, this result is quite tight. $\frac{\pi}{8\sqrt{3}} \approx 0.227$, and it is pretty easy to show that the best possible bound is at most 0.25.

12. Suppose the convex hull of the points has k points on the boundary, perimeter P, and area A. Prove that some triangle has area at most $\min\left(\frac{P^2}{2k^2}, \frac{A}{2(n+1)^2-k-2}\right)$.

13. Fix a permutation σ of $\{1, 2, \ldots, n\}$, and consider the sets $X_i = \{\sigma(i), \sigma(i+1), \ldots, \sigma(i+k-1)\}$ (taking all indices mod k).

14. Fix a permutation σ of $\{1, 2, \ldots, n\}$, and consider the sets $X_i = \{\sigma(1), \sigma(2), \ldots, \sigma(i)\}$. Calculate the expected number of sets X_i in $\mathcal{C}$.

15. Pick a random labeling $\{1, 2, \ldots, n\}$ of X. Consider the event E_i where every element in A_i has label less than every element in B_i. Use the fact that $1 \geq$ Prob[At least one event E_i happens].

16. First show the number of crossing edges is at least $E - 3V$ (using the fact that a planar graph has $V - E + F = 1 \implies E < 3V$). Now sample vertices independently with some probability p and apply this inequality.

17. Consider the following algorithm for finding a cut. Choose a random edge. Merge the two endpoints, deleting all edges between those endpoints but keeping all other edges. Note that it may be possible to now have multiple edges between the same two vertices – that's ok. Repeat this contraction process until only two vertices remain. The remaining edges form a cut of the original graph (why?). Given a fixed min-cut, what is the probability that this algorithm will find it at the end?