How To Prove It

SHAILESH SHIRALI

The topic of 'proof by induction' is now a standard part of the syllabus of mathematics at the 11-12 level. Most students consider it a 'scoring topic' – they generally master the mechanics of proof by induction quickly, as the proofs follow a standard trajectory and are easy to mimic.

My experience as a mathematics teacher, however, suggests that the vast majority of students do not grasp what proof by induction is all about. While they are able to mimic all the required steps, most of them do not grasp the essential logic of such proofs. Indeed, to a good many students, these proofs give the impression of circular reasoning! For someone steeped in the culture of mathematics, it is not easy to understand why students find it so difficult to grasp the essence of such proofs. Is it because the topic is taught in haste, with insufficient time spent on the subtleties involved (and there surely are many subtleties involved)? Or is it because proof itself is inherently a difficult topic? I suppose that a great deal more research is needed to understand the core of the difficulty. It would be well worth it for teachers themselves to undertake such research, rather than wait for experts to take up the task.

In this and the following episode of *How to Prove It*, we shall dwell on some critical aspects of induction proofs (aspects which possibly are not emphasised strongly enough) and study a few examples that show how valuable and versatile it is as a proof technique.

Structure of a proof by mathematical induction

We start by listing the essential components of a proof by induction:

(1) Framing the hypothesis or conjecture.

(2) Anchoring the induction, i.e., verifying the initial step.

(3) The bridge step, i.e., establishing the link between successive propositions of the induction hypothesis.

We now elaborate on each of these steps.

Keywords: Proof by induction

Framing the hypothesis or conjecture

This step is generally not encountered at the school level; the student is simply *given* the hypothesis or conjecture to be proved and asked to prove it using the principle of induction (i.e., asked only to execute the second and third steps). So the student is simply being asked to provide the algebraic details of the proof. *But this step is crucial.* Indeed, it is this step that constitutes what is 'inductive' about proof by induction. Without it, including the word 'induction' in 'proof by induction' is meaningless; the exercise cannot be called a proof by induction. For, if we skip over this initial step, a proof by induction is actually a proof by deduction! So it is most unfortunate that this step is completely absent in the way the topic is taught in schools at present.

'Framing the hypothesis' can be expressed more simply as (and most often reduces to): *Guess the formula!* Accordingly, we shall study a few examples to illustrate this.

Example 1: Partial sums of the consecutive integers. Let's start with a simple and familiar example: guessing a formula for the sequence of partial sums of the consecutive positive integers. That is, we seek a formula for the n-th term of the sequence

$$1, \quad 1+2, \quad 1+2+3, \quad 1+2+3+4, \quad 1+2+3+4+5, \quad \dots. \tag{1}$$

Performing the additions, we obtain the following sequence:

$$1,\ 3,\ 6,\ 10,\ 15,\ 21,\ 28,\ 36,\ 45,\ 55,\ \dots \tag{2}$$

How do we proceed now?

There is no standard way to guess a formula for a given sequence. Rather, we have to play with the sequence and hope for the best! That is, we may double all the terms; or multiply all the terms by some other number; or add some constant to all the terms; or add and subtract some constant to the terms in alternation; or divide the terms by their serial number; or express the terms in factorized form; and so on. In short, we have to perform all kinds of arithmetical operations to the given sequence, in the hope of uncovering some visible pattern, one which we can 'catch hold of' and which will enable us to formulate a suitable hypothesis or conjecture. *It is this which is the inductive step*—i.e., guessing a formula or formulating a conjecture.

In this instance, doubling all the terms yields:

$$2,\ 6,\ 12,\ 20,\ 30,\ 42,\ 56,\ 72,\ 90,\ 110,\ \dots \tag{3}$$

Now, if we divide each term by its position number (i.e., the position where it occurs in the sequence), that is, we do $2 \div 1$, $6 \div 2$, $12 \div 3$, $20 \div 4$, …, then the pattern is instantly revealed. We get:

$$2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ \dots \tag{4}$$

The n-th term of this sequence is 'clearly' $n+1$, which means that the n-th term of the sequence $2, 6, 12, 20, 30, \dots$ must be $n(n+1)$, and therefore the n-th term of the original sequence $1, 3, 6, 10, 15, \dots$ must be $\frac{1}{2}n(n+1)$. This is what inductive thinking has led us to believe; it represents our educated guess for a formula giving the sum of the first n positive integers. At this stage, it is, of course, only a guess; we have not proved anything as yet.

Example 2: Partial sums of the squares of the consecutive integers. We consider now a more complex example: guessing a formula for the sequence of partial sums of the squares of the consecutive positive integers. That is, we seek a formula for the n-th term of the sequence

$$1^2, \quad 1^2+2^2, \quad 1^2+2^2+3^2, \quad 1^2+2^2+3^2+4^2, \quad 1^2+2^2+3^2+4^2+5^2, \quad \dots \tag{5}$$

Performing the additions, we obtain the following sequence:

$$1,\ 5,\ 14,\ 30,\ 55,\ 91,\ 140,\ 204,\ 285,\ 385,\ \ldots \tag{6}$$

As the sequence grows more rapidly than the earlier one, we should be prepared to do more experimentation before we can guess the formula. Let's try 'cutting the numbers down to size,' by dividing each term by its position number. That is, if s_n denotes the sum $1^2 + 2^2 + \cdots + n^2$, then let's compute the values of $s_n \div n$. We get:

$$1,\ \frac{5}{2},\ \frac{14}{3},\ \frac{15}{2},\ 11,\ \frac{91}{6},\ 20,\ \frac{51}{2},\ \frac{95}{3},\ \frac{77}{2},\ \ldots \tag{7}$$

We observe quickly the repeated occurrence of the denominators 2, 3, 6. Continuing the computations, we find that the behaviour does not change; no other denominator turns up. (In passing, we note that the sequence of denominators has a nice pattern: 1, 2, 3, 2, 1, 6, 1, 2, 3, 2, 1, 6, and so on, with the string $1, 2, 3, 2, 1, 6$ repeating indefinitely.) This prompts us to multiply the latest sequence by the LCM of $2, 3, 6$, i.e., by 6. (This kind of ad hoc logic is typical of the inductive stage. We have to be prepared to do all kinds of computations. Most of the time, these efforts do not lead to anything at all.) Here's what we get:

$$6,\ 15,\ 28,\ 45,\ 66,\ 91,\ 120,\ 153,\ 190,\ 231,\ \ldots \tag{8}$$

These are the values of $6s_n \div n$. A pattern is quickly noticed when we express the numbers in factorized form:

$$2 \times 3,\ 3 \times 5,\ 4 \times 7,\ 5 \times 9,\ 6 \times 11,\ 7 \times 13,\ 8 \times 15,\ 9 \times 17,\ 10 \times 19,\ 11 \times 21,\ \ldots \tag{9}$$

It does not take much effort to guess that this is the sequence $(n + 1)(2n + 1)$. This means that $6s_n \div n = (n + 1)(2n + 1)$, i.e.,

$$s_n = \frac{n(n+1)(2n+1)}{6}. \tag{10}$$

This is what inductive thinking has led us to believe; it represents our educated guess for a formula giving the sum of the squares of the first n positive integers. As earlier, it is only a guess; we have not proved anything as yet.

Example 3: Spearman's rank correlation coefficient: an associated calculation. In the study of rank correlation, we wish to define a measure indicating how close two sets of ranks are to each other. Mathematically, it amounts to finding a measure of closeness of two different permutations of the string $(1, 2, \ldots, n-1, n)$, where n is any positive integer. (To start with, let us not consider the possibility of tied ranks, which naturally introduces complications.) We would like the measure to be such that it assumes a value of $+1$ when the permutations are identical to each other, and a value of -1 when the permutations are reverses of each other.

Let the two permutations of $(1, 2, \ldots, n-1, n)$ be

$$(a_1, a_2, \ldots, a_n), \qquad (b_1, b_2, \ldots, b_n) \tag{11}$$

respectively. A natural measure of how different they are from each other is the following score S:

$$S = (a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2. \tag{12}$$

Note that we have opted to square the differences between respective ranks to avoid the situation where negative values cancel out positive values. (Instead of taking the squares, we could also have taken the absolute values. This would be perfectly acceptable.)

How do we convert the score S into a coefficient that lies between -1 and 1? Here's a simple and elegant way of doing this. The minimum possible value of S is clearly 0; this is attained when the two permutations are identical to each other. What is the maximum possible value of S? Denote this maximum value by M. Now compute the quantity

$$\rho = 1 - \frac{2S}{M}. \tag{13}$$

Observe that if the two permutations are identical to each other, then $S = 0$, leading to $\rho = 1$, which is as it should be; and if the two permutations are as far apart as possible from each other, then $S = M$, leading to $\rho = -1$, which once again is as it should be. So this formula makes sense. It remains now to find a convenient formula for M.

Let us compute the value of M for various values of n. This will be the value of the score S when the two permutations are the following:

$$(1, 2, \ldots, n-1, n), \qquad (n, n-1, \ldots, 2, 1). \tag{14}$$

That is,

$$M = (1-n)^2 + (2-n+1)^2 + \cdots + (n-1-2)^2 + (n-1)^2. \tag{15}$$

Let us now prepare a table of values of M (we could use a computer to do this, if we wish). Here is what we get.

n	1	2	3	4	5	6	7	8	9	10	$\cdots$
M	0	2	8	20	40	70	112	168	240	330	$\cdots$

(16)

Let us now try to find a formula for the sequence

$$0,\ 2,\ 8,\ 20,\ 40,\ 70,\ 112,\ 168,\ 240,\ 330,\ \ldots \tag{17}$$

Denote the n-th term of the sequence by s_n. Observe that the sequence is growing quite rapidly. So let us produce a new sequence defined by $s_n \div n$. We obtain the following:

$$0,\ 1,\ \frac{8}{3},\ 5,\ 8,\ \frac{35}{3},\ 16,\ 21,\ \frac{80}{3},\ 33,\ \ldots \tag{18}$$

We see that a denominator of 3 occurs at regular intervals in this sequence. So let us multiply the above sequence by 3, to clear the denominators. We now obtain the following:

$$0,\ 3,\ 8,\ 15,\ 24,\ 35,\ 48,\ 63,\ 80,\ 99,\ \ldots \tag{19}$$

We could now experiment further with this sequence, but recognition is faster! We notice easily enough that each number in this sequence is 1 less than a perfect square. Indeed, if we add 1 to each number in the sequence, we simply obtain the sequence of perfect squares. It follows that the n-th term of (19) is $n^2 - 1$, hence the n-th term of (18) is $\left(n^2 - 1\right) \div 3$, hence the n-th term of (17) is $\left(n^3 - n\right) \div 3$. Hence:

$$s_n = \frac{n^3 - n}{3}. \tag{20}$$

So we have obtained a formula for the coefficient of rank correlation (defined as above):

$$\rho = 1 - \frac{6S}{n^3 - n}. \tag{21}$$

By following this path, we have hit upon the formula for an extremely famous measure of rank correlation: ***Spearman's coefficient of rank correlation***. See [2].

Important comment. It is important to note exactly what we have accomplished till now. We have been able to find formulas for the general term of three important sequences, using a mixture of play and experimentation and guesswork. So *we have completed the inductive part of the exercise*: i.e., we have been able to frame conjectures that fit the data.

It now remains to prove each of the conjectures. It is at this stage that what is generally called 'proof by induction' enters the picture. So we have to: (i) anchor the induction (i.e., verify the initial part of the conjecture); (ii) prove the bridge step (i.e., verify the link between successive propositions of the induction hypothesis).

We will elaborate on the latter two steps in a follow-up article.

References

1. Wikipedia, "Mathematical induction" from https://en.wikipedia.org/wiki/Mathematical_induction
2. Wikipedia, "Spearman's rank correlation coefficient" from https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

SHAILESH SHIRALI is Director of Sahyadri School (KFI), Pune, and Head of the Community Mathematics Centre in Rishi Valley School (AP). He has been closely involved with the Math Olympiad movement in India. He is the author of many mathematics books for high school students, and serves as Chief Editor for *At Right Angles*. He may be contacted at shailesh.shirali@gmail.com.