Expected trials under the matching rounds problem

Jim Farmer
Macquarie University
<jim.farmer@mq.edu.au>

In which an argument with religious fundamentalists inspires an elegant expected number of trials exercise based on a generalisation of the *matching problem*.

Review of the matching problem

We begin by reviewing the famous matching problem, since its result will prove useful in the harder problem we will subsequently consider.

We have n objects numbered 1 to n, and n similarly numbered cells, $n \in \{1, 2, 3, \ldots\}$. We intend to randomly place the objects into the cells, subject to the constraint that exactly one object is placed in each cell. We regard an object as 'correctly placed' if it is placed in the cell of the same number.

Let L be a random variable denoting the number of correctly placed objects.

Depending on the source, the 'matching problem' may refer to the calculation of any or all of the following:

- the probability that $L = 0$, which can be found by the Principle of Inclusion and Exclusion;
- the probability function of L, which requires a generalisation of the Principle of Inclusion and Exclusion;
- the expected value of L, which is easily found using indicator random variables;
- the variance of L.

In this article we are only concerned with the third problem, the expected number of objects that are correctly placed.

The matching problem is often presented in more concrete guises. Here are three versions suitable for the classroom. The first two are common in older probability texts, but appear somewhat dated.

- *The confused secretary:* A secretary has typed n letters, and has also typed the appropriate addresses onto n envelopes. The letters and envelopes are passed to a second secretary with the instruction: "Put the letters in

the envelopes." Due to the instruction containing insufficient detail, the second secretary mistakenly assumes the letters are all identical and places them at random into the envelopes, one letter per envelope. What is the expected number of correctly placed letters?

- *The ballroom party: n* heterosexual married couples hold a dance party. Each male randomly chooses one female for the first waltz, but cannot choose a female who has already been chosen. What is the expected number of men who chose their own wife?
- *The secret Santa: n* workmates place their names in a bucket. Each person chooses a name at random from the bucket and then must buy a jovial gift for that person to be presented anonymously at the office end of year party. What is the expected number of people who are relieved to find they chose their own name?

The solution to the matching problem

Let L_i be an indicator random variable denoting the number of times that object i is placed in its correct cell, $i = 1, 2, 3, \ldots, n$. That is, L_i takes the value 1 if object i is placed in cell i and 0 if it is not.

The number of ways to arrange the n objects in the cells, one object per cell, is $n!$. The number of arrangements that have the ith object correctly placed is $(n-1)!$, since the other $n-1$ objects can be placed in the other $n-1$ cells, one object per cell, in any order.

$$P(L_i = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

$$E(L_i) = 0 \times P(L_i = 0) + 1 \times P(L_i = 1)1 = P(L_i = 1) = \frac{1}{n}$$

$$E(L) = E\left(\sum_{i=1}^{n} L_i\right) = \sum_{i=1}^{n} E(L_i) = \sum_{i=1}^{n} \frac{1}{n} = 1$$

That is, we arrive at the pleasant result that the expected number of correctly placed objects is 1, a result that is independent of the size of n.

It is instructive to note that the L_i random variables are not independent. The above proof falls out very simply because the property

$$E\left(\sum_{i=1}^{n} L_i\right) = \sum_{i=1}^{n} E(L_i)$$

does not require independence. By contrast, the similar property for variances does require independence, so the task of finding $VAR(L)$ is much harder, though the final answer is again surprisingly simple.

The extension: The matching rounds problem

As above, we have n objects labelled 1 to n and n corresponding cells. Arrange the objects randomly in the cells, one object per cell. Identify any objects that were correctly placed, meaning they were placed in the similarly numbered cell. Discard these objects and their cells. They play no further part in the experiment. This concludes the first trial. If all the objects were correctly placed then the experiment has concluded in a single trial.

If any objects were not correctly placed on the first trial, perform a second trial. The remaining objects—meaning those not correctly placed on the first trial—are randomly arranged in the remaining cells, one object per cell. Identify any objects correctly placed on this second trial. These objects and their cells are discarded and take no further part in the experiment. The remaining objects and cells, if any, progress to the third trial.

We continue these trials until all objects have been correctly placed. We seek the expected number of trials.

Investigating small values of *n*

Let $f(n)$ denote the expected number of trials if we start with n objects and n cells.

The $n = 1$ case is trivial. If there is only 1 object it must be correctly placed on the first trial. Hence $f(1) = 1$.

When $n = 2$ there are two possible equally likely outcomes for the first trial. It is not possible for exactly one object to be correctly placed. Either both objects are correctly placed, and the experiment concludes in one trial, or they are both incorrectly placed, and we have made no progress. In the latter case we have completed 1 trial, but since we made no progress the expected number of further trials required is still $f(2)$. Hence the conditional expectation theorem gives:

$$f(2) = \frac{1}{2} \times 1 + \frac{1}{2} \times \big(1 + f(2)\big)$$

For readers not familiar with the conditional expectation theorem, this result could also be written in the following form, which seems intuitively reasonable:

$$f(2) = 1 + \frac{1}{2} \times 0 + \frac{1}{2} f(2)$$

That is, we definitely must make the first trial so we count that on the right hand. There is a chance of $\frac{1}{2}$ that this trial completes the experiment, so no further trials are required, and a chance of $\frac{1}{2}$ that both objects are incorrectly placed, so we have made no progress and the expected number of further trials is $f(2)$.

Whichever form is used, the equation easily solves to $f(2) = 2$.

As an aside, when $n = 2$, the number of trials is a geometric random variable with parameter $\frac{1}{2}$. Since the expected value of a geometric random variable is the reciprocal of its parameter, $f(2) = 2$. However, this insight is of no help when $n > 2$.

When $n = 3$ it is still simple enough to list the 3! outcomes. We can imagine arranging the three objects at random in a line, and placing the leftmost object in cell 1, the middle object in cell 2 and the rightmost object in cell 3. The possible orders of the objects are:

- (1, 2, 3). All objects are correctly placed on the first trial.
- (1, 3, 2), (3, 2, 1) and (2, 1, 3). Exactly one object is correctly placed on the first trial, so the second trial will involve two objects, meaning the expected number of further trials required is $f(2)$.
- (2, 3, 1) and (3, 1, 2). All objects are incorrectly placed on the first trial, so we have made no progress and the second trial will involve three objects, so the expected number of further trials required is $f(3)$.
- It is not possible for exactly two objects to be correctly placed.

Hence the conditional expectation theorem gives

$$f(3) = \frac{1}{6} \times 1 + \frac{3}{6} \times \big(1 + f(2)\big) + \frac{2}{6} \times \big(1 + f(3)\big)$$

Alternatively, the more intuitive form gives

$$f(3) = 1 + \frac{1}{6} \times 0 + \frac{3}{6} \times f(2) + \frac{2}{6} \times f(3)$$

Since $f(2) = 2$, either form is easily solved giving $f(3) = 3$.

It is also simple to list the 4! = 24 possible outcomes for the first trial when $n = 4$ and derive the result $f(4) = 4$. Hence we suspect that in general $f(n) = n$, a result which we can prove using the strong form of mathematical induction.

The general result by induction

The result $f(1) = 1$ gives a starting point for the induction process.

Assume $f(n) = n$ for $n = 1, 2, 3, \ldots, k - 1$.

The induction step requires us to prove that $f(k) = k$.

Let $g_k(t)$ denote the probability that, on the first trial of an experiment involving k objects and cells, exactly t objects are placed in their correct cells.

Exact expressions for $g_k(t)$ can be developed using a generalisation of the Principle of Inclusion and Exclusion, and I have happily spent many hours and pages doing so before realising it was unnecessary. Surprisingly, we only require the following two identities:

$$g_k(0) + g_k(1) + g_k(2) + \ldots + g_k(k) = 1 \qquad (1)$$

$$g_k(1) + 2g_k(2) + 3g_k(3) + \ldots + kg_k(k) = 1 \qquad (2)$$

Equation 1 is true because the left-hand side sums the probabilities for mutually exclusive and exhaustive events. That is, if we randomly place k numbered objects in k correspondingly numbered cells, one object per cell, the number of objects which are correctly placed must be a number in the set $\{0, 1, 2, \ldots, k\}$. While it is not possible for exactly $k-1$ objects to be correctly placed, including the zero probability $g_k(k-1)$ in the left hand side does no harm.

Equation 2 is true because the left-hand side is an expression for the expected number of objects correctly placed on the first trial. This is 1, the answer to the matching problem considered earlier.

The conditional expectation theorem gives

$$f(k) = g_k(0)\{1 + f(k)\} + g_k(1)\{1 + f(k-1)\} + g_k(2)\{1 + f(k-2)\} + \ldots \\ + g_k(k-1)\{1 + f(1)\} + g_k(k)\{1\}$$

Employing equation (1) simplifies this to

$f(k) = 1 + g_k(0)f(k) + g_k(1)f(k-1) + g_k(2)f(k-2) + \ldots + g_k(k-1)f(1) + g_k(k) \times 0$

which is the more intuitive form we can justify directly as follows. Looking at the terms on the right hand side:

- The 1st term says that the first trial definitely happens.
- The 2nd term says that there is a probability of $g_k(0)$ that none of the k objects are correctly placed on the first trial, in which case we have made no progress and the expected number of further trials required is $f(k)$.
- The 3rd term says that there is a probability of $g_k(1)$ that exactly one of the k objects is correctly placed on the first trial, in which case the second trial will involve $k-1$ objects, so the expected number of further trials required is $f(k-1)$.

  ⋮
- The last term states that there is a probability of $g_k(k)$ that all k objects are correctly placed on the first trial, in which case the number of further trials required is zero.

The induction assumption simplifies this to:

$$f(k) = 1 + g_k(0)f(k) + (k-1)g_k(1) + (k-2)g_k(2) + \ldots \\ + 1 \times g_k(k-1) + 0 \times g_k(k)$$

$$\{1 - g_k(0)\}f(k) = 1 + k\left[g_k(1) + g_k(2) + \ldots + g_k(k-1) + g_k(k)\right] \\ - \left[g_k(1) + 2g_k(2) + \ldots + (k-1)g_k(k-1) + kg_k(k)\right]$$

Use equations (1) and (2) to simplify the two expressions in square brackets on the right hand side.

$$\{1 - g_k(0)\}f(k) = 1 + k\left[1 - g_k(0)\right] - 1$$

$$f(k) = k$$

Hence, by the strong form of mathematical induction, we can conclude $f(n) = n$, $n = 1, 2, 3\ldots$

Other methods

Of course, the delight of discovering a solution to a new interesting problem is often followed by the disappointment of finding it has already been done. Ross (1972) calls this problem the matching rounds problem, though this name does not appear to be in common use, and solves it in Example 3.13 by the method I have given above.

This appears to be the easiest solution, but those who like challenging combinatorics manipulations might like to refer to a solution by Griffiths (2002). Griffiths also identifies other published methods, though these tend to require knowledge of more advanced techniques that may feel far too complex relative to the difficulty of the problem being solved.

Motivation

Some religious fundamentalists interpret religious documents literally and so believe that the universe and all life we see on earth was created about 6000 years ago by a supernatural being. Most of these people choose simply to ignore the scientific evidence that our planet is about 4.5 billion years old and that life has been evolving for a large portion of that period. However, some attempt to dispute the scientific evidence and may present arguments which either misunderstand or misrepresent that evidence.

The TalkOrigins Archive is a website that grew out of the talk.origins Usenet group. The website's introduction states: "The primary reason for this archive's existence is to provide mainstream scientific responses to the many frequently asked questions (FAQs) that appear in the talk.origins newsgroup and the frequently rebutted assertions of those advocating intelligent design or other creationist pseudosciences" (TalkOrigins Archive, 2006).

Since the majority of errors made by religious opponents of evolution fall into a few simple categories, the TalkOrigin Archive includes "Five Major Misconceptions about Evolution" (Isaak, 2003). The fourth misconception is "The theory of evolution says that life originated, and evolution proceeds, by random chance."

Evolution does not deal with how life originated. That falls under the heading of abiogenesis. Evolution deals with how living things change over time. Evolution is not a purely random process, but rather requires both random genetic mutation and non-random natural selection.

Opponents of evolution often lose the natural selection component and seek to portray evolution as an entirely random process, and then argue that

the chance of something as complex as a human arising from an entirely random process is too small to consider.

For example, they may quote Morris' (1974) claim that "Complex structures could not have arisen by chance", which has been dealt with by Isaak (2004a, 2005).

They may quote Hoyle's (1983) inappropriate (Isaak, 2004b) analogy that: "Order does not spontaneously form from disorder. A tornado passing through a junkyard would never assemble a 747."

That is, they seem to believe that evolution requires an entire organism to be assembled in a single random act rather than evolving in gradual steps that combine random mutation and non-random selection. (If that were true, scientists would not call it 'evolution'; they would name it something like 'spontaneous formation'.)

Critics of evolution have used other similar invalid analogies, with Hoyle's 747 replaced by some other complex structure. One that I have heard several times but have not located in print is: "If you shake a jigsaw puzzle box, you wouldn't expect to open the box and find the shaking had magically assembled the puzzle." Of course you wouldn't. For a start, the solved puzzle usually significantly exceeds the dimensions of the box, so it would not fit. But more importantly this argument only models the random mutation part (and not very well) and loses the non-random natural selection. While talking with proponents of this analogy, I began to get the feeling that some of the proponents did understand that evolution requires non-random selection, but that they felt this was not a strong enough force to impact significantly on the random mutations. This prompted me to develop a webpage containing some simple animations based on the jigsaw puzzle analogy (Farmer, 2014).

I began with a square image of the young Charles Darwin which I divided into four smaller squares, giving a crude four-piece jigsaw puzzle. As a demonstration of random mutation alone, the four pieces are placed at random onto a two by two grid, one piece per cell. In this simple model, rotation of the pieces is not allowed. The result is examined to see whether it has reformed the original image. If it has not, the pieces are all removed, and the process is repeated until success is achieved.

A second animation adds a non-random element that mimics selection. The four pieces are again placed at random. If all pieces are correctly placed, the process stops. If not, any pieces placed in their correct cell are left in place —modelling a non-random selection effect—with only the misplaced pieces being removed. The misplaced pieces are then randomly placed in any empty cells, and the process is repeated until all pieces are correctly placed.

The two animations run side by side, so the viewer can see the second animation usually completing the task more quickly.

An obvious question is: for each animation, what is the expected numbers of trials to completion? The fact that the four pieces are being placed in a

square arrangement is not relevant. We can simply think of the pieces and the grid as four objects and four cells, each object having a 'correct' cell.

The first animation is straightforward. There are 4! or 24 possible arrangements. Since a failure results in all four pieces being removed, the number of trials is a geometric random variable with parameter $\frac{1}{24}$, and hence the expected number of trials, being the reciprocal of the parameter, is 24.

The second animation is the matching rounds problem described in this paper with $n = 4$, and hence the expected number of trials is 4.

To give some idea how quickly these results change as the complexity of the problem increases, the animation can be adjusted to increase the number of pieces in the jigsaw puzzle from 4 to 16. In the random placement animation, the expected number of trials to completion is $16! \approx 2 \times 10^{13}$. Not surprisingly, I have never seen this animation complete. When the selection process is added, allowing correctly placed pieces to stay in place with only incorrectly placed pieces being removed, the expected number of trials drops to 16. Perhaps animations like these can give the less mathematical some inkling of the power of natural selection.

I should state clearly that I do not regard these animations as an accurate model of how evolution really works. The first animation, which removes all the pieces and starts again if success is not achieved, is clearly an atrocious model, since it does not even attempt to include a natural selection component. The second animation, which models natural selection by allowing correctly placed pieces to stay in place, is considerably less atrocious, but is still a long way from reality. For example, it has only one solution, the reforming of the original image, whereas in reality the set of viable life forms is huge, probably much larger than the set of life forms that have ever existed. Biologists have certainly invented much better analogies than the second animation, but that need not stop maths-oriented life forms from appreciating the elegance of the solution to the expected value problem it spawned.

Relevance to the Australian Curriculum

If trying to present the ideas in this paper to advanced Year 12 mathematics students, the challenging part is the conditional expectation theorem, which students will not usually encounter formally until a tertiary probability course. However, we only require the discrete random variable form of the conditional expectation theorem, and this form does feel intuitively reasonable. Some describe it as working out an expected value using a two-step tree diagram. When students use a two-step tree diagram to evaluate a probability, they are really using the partition theorem, and the conditional expectation theorem does a very similar thing to evaluate an expected value. In this paper, when the conditional expectation theorem was required, care was taken to also present

the argument in a form that hopefully feels intuitively sensible to those not familiar with that theorem.

In terms of the senior mathematics courses in the *Australian Curriculum*, the other major skills used in this paper fall in the Specialist Mathematics and Mathematical Methods courses.

The basic form of mathematical induction is dealt with in the Specialist Mathematics course. This paper uses the strong form of mathematical induction. A student with a good understanding of the basic form will hopefully find the strong form convincing.

The Mathematical Methods course provides the other necessary skills in probability, random variables, expected values and combinatorics. While this paper refers to indicator random variables, which are not explicitly mentioned in the syllabus, these are merely a special case of a random variable and should not cause difficulties.

The Essential Mathematics course includes "perform simulations of experiments using technology". The animations referred to in this paper arguably fall in this category, but the problem considered here is probably more difficult than intended by the curriculum.

Evolution appears in the Year 10 science syllabus, and is developed further in the senior biology course. While the mathematics in this paper is well beyond Year 10 level, perhaps the animations referred to may assist Year 10 science students interested in the pseudoscience sometimes employed by those trying to discredit evolution.

References

Farmer, J. A. (2014). *The evolving jigsaw puzzle*. Retrieved January 2014 from http://math.plussed.net/evolution_jigsaws/index.php

Griffiths, M. (2002). The 'self santa' problem. *The Mathematical Gazette, 86*(507), 487–489. Retrieved January 2014 from http://www.jstor.org/stable/3621154.

Hoyle, F. (1983). *The intelligent universe*. New York: Holt, Rinehart & Winston.

Isaak, M. (2003). F*ive major misconceptions about evolution*. Retrieved January 2014 from http://www.talkorigins.org/faqs/faq-misconceptions.html

Isaak, M. (2004a). *CB940: Complex structures by chance*. Retrieved January 2014 from http://www.talkorigins.org/indexcc/CB/CB940.html

Isaak, M. (2004b). *CF002.1: Tornadoes in a junkyard*. Retrieved January 2014 from http://www.talkorigins.org/indexcc/CF/CF002_1.html

Isaak, M. (2005). *CB940.1: The mathematical probability of evolution*. Retrieved January 2014 from http://www.talkorigins.org/indexcc/CB/CB940_1.html

Morris, H. M. (1974). *Scientific creationism*. Green Forest, AR: Master Books.

Ross, S. (1972). *Introduction to probability models* (6th ed.). San Diego, CA: Academic Press.

TalkOrigins Archive. (2006). *Exploring the creation/evolution controversy*. Retrieved January 2014 from http://www.talkorigins.org