Stickers

There's set of 221 stickers and you want to collect them all and every time you buy a piece of bubble gum you get one of these at random. How many pieces of gum do you have to buy to get all 221 of them?

We start by asking what the problem means. What is the force of those words "on average"? Well it's this: suppose there were a large number of children, all (independently) trying to complete the set. Now you expect that the total number of stickers each child will have to buy will vary (probably by quite a bit)—some will be luckier than others. But if you took the average of that total number over all these children, you'd get an estimate of the (theoretical) answer we're looking for here.

Some simpler problems

It seems like an impossibly hard problem, so we work up to it slowly. It would be easier if there were a smaller set of possible stickers. A two-sticker book can be modeled with a coin—how many tosses on average are required to get both a head and a tail? And a six-sticker book can be modeled with a die— what is the average number of rolls required to get all six outcomes?

Collecting some data

There were forty students and I gave each of them a die and asked them to roll as many times as needed to get every number at least once, and record the number of rolls required. And then to do it again.

Out of the 80 experiments, 2 managed to need only 6 rolls (a new number each time!), 1 got it in 7 rolls, 4 got it in 8 rolls, etc. The results are tabulated below.

Number of rolls to get every number at least once

# Rolls	Occurrences	# Rolls	Occurrences
6	2	16	5
7	1	17	1
8	4	19	3
9	3	21	1
10	6	22	1
11	8	25	1
12	10	27	1
13	11	32	1
14	9	33	1
15	10	40	1
		Total	80

I begin by displaying the Panini sports book my son, age 7, had managed to completely fill during my sabbatical year in England. It contained spaces for the pictures of 221 athletes of different sports, and by the end of the year, all 221 had been filled. What happens is that you buy these stickers, five in a package (along with bubble gum), but there's no way to know which pictures you get. I remember James' delight at the beginning, in the fall, when he'd always get 3 or 4 new ones in a package, and often all 5 would be new. But by the time the damp grey Oxford January set in, he was lucky to get 1 or 2 new ones in a pack; often all five would be repeats. It occurred to him that the last few stickers were going to be very hard to get. Luckily, about this time he found a friend at school who was also working on the same book, and the ability to swap duplicates made the last part of the journey much more fun and more rewarding. The job was done before the April violets laid their magic carpet of blue throughout the Oxford woods.

This empirical phase of the problem is important. It gets them active and interacting, and it also gives them a feeling for the large number of different possible “scores” that the average represents. And I find the 6-sticker problem is just about the right size for the classroom—and they do love rolling those dice

The sum of the two “occurrences” columns is the total number of experiments which is 80.

Now the average # of rolls required is found by adding up all the outcomes we obtained and dividing by the total number of experiments. That is, we have to add two 6's, and one 7, and four 8's, etc. and divide by 80:

$$\text{Avge \# rolls} = \frac{2\cdot 6 + 1\cdot 7 + 4\cdot 8 + \cdots + 1\cdot 40}{80} = \frac{1119}{80} \approx 14.0$$

Note that the numerator really is the sum of 80 numbers, but I've grouped the 2 6's together, etc. We get 14.0 as an empirical estimate of the theoretically exact value of the average.

If you like you can regard the multipliers 2, 1, 4, etc. as weights, and the 80 in the denominator is the sum of the weights. That's what makes it an average.

Attacking the theoretical average.
I pose the following graded sequence of questions.

1) I have a fair coin. What is the average number of tosses to get the first head?

The answer is 2. In fact this is an immediate application of the dart board theorem, which I have just investigated with this class. With each toss, I get a head with probability ½. That's just like having probability ½ of hitting the bulls-eye.

I could of course have kept the big problem on the table and let the class try to construct this sequence of questions leading up to it themselves. And sometimes that's what I would do. But sometimes, too, they like to be pampered. Also, the simplicity of the questions provided an opportunity for a number of students to come to the board with solutions

2) I have a fair coin. What is the average number of tosses to get both a head and a tail?

The answer is 3. On the first toss I get a head or a tail. Either way, to get the remaining outcome is a problem 1) situation and will take on average 2 tosses—for a total of 3.

3) I have a fair die. What is the average number of tosses to get a **1**?

The answer is 6. This is the dart board principle (**darts**) with probability 1/6 of hitting the bulls-eye.

4) I have a fair die. What is the average number of tosses to get both a **1** and a **2**?

The answer is 9. It takes a moment to see how to set this one up. We first ask how long it takes to get either a **1** or a **2**. Well, these two outcomes represent 1/3 of the possibilities, so this is the dart board principle again with probability 1/3 of hitting the bulls-eye, and it will take an average of 3 tosses to get one or the other. Then having done that, we are in a problem 3) situation and it will take 6 more tosses to get the other.

What we are actually developing here is an extremely elegant and powerful method of analysis, using the dartboard principle again and again.

5) I have a fair die. What is the average number of tosses to get a **1**, a **2** and a **3**?

The answer is 11. First ask how long it takes to get one of the three. Well, these three outcomes represent 1/2 of the possibilities, so by the dart board principle, we expect to take an average of 2 tosses. Then having done that, we are in a problem 4) situation and it will take 9 more tosses to get the rest.

6) I have a fair die. What is the average number of tosses to get a **1**, a **2**, a **3**, and a **4**?

The answer is 12.5. First ask how long it takes to get one of the four. Well, these four outcomes represent 2/3 of the possibilities, so by the dart board principle, we expect to take an average of 3/2 tosses. Then having done that, we are in a problem 5) situation and it will take 11 more tosses to get the rest.

7) Time to jump to the end. I have a fair die. What is the average number of tosses to get each outcome at least once?

On the first toss we get one of the six. The remaining five outcomes represent 5/6 of the possibilities, so it will take an average of 6/5 tosses to get the next. Then having done that, we are in a problem 6) situation and it will take 12.5 more tosses to get the rest. The total is 1+1.2+12.5 = 14.7.

We see that the empirical estimate 14.0 that we obtained was not far off—it was "trying" to be 14.7.

Now if we want to generalize this result to find the average number of stickers required to fill the *Panini* book, we want to replace 6 by 221, and to do this we need an "anatomy" of that number 14.7.

Here's how to think of it. The first "success" came on the first roll, so took 1 roll. The second success was a dart board problem with probability 5/6 so took 6/5 rolls. The next success was a dart board problem with probability 4/6 so took 6/4 rolls. The next success was a dart board problem with probability 3/6 so took 6/3 rolls. The next success was a dart board problem with probability 2/6 so took 6/2 rolls. Etc. Thus:

$$\text{Average \# rolls} = 1+\frac{6}{5}+\frac{6}{4}+\frac{6}{3}+\frac{6}{2}+\frac{6}{1} = 14.7.$$

To see how to generalize the dice result, we need to understand exactly how the answer depends on the number 6. How is each piece of the answer related to 6?

There's a simple pattern here. That initial 1 was actually 6/6, and the total can be written

$$\text{Avge \# rolls} = \frac{6}{6}+\frac{6}{5}+\frac{6}{4}+\frac{6}{3}+\frac{6}{2}+\frac{6}{1}$$

$$= 6\left[\frac{1}{6}+\frac{1}{5}+\frac{1}{4}+\frac{1}{3}+\frac{1}{2}+\frac{1}{1}\right] = 14.7$$

What a fine formula!

And now we can solve the sticker problem—replace 6 by 221. The avge # stickers is:

$$221\left[\frac{1}{221}+\frac{1}{220}+\frac{1}{219}+\frac{1}{218}+\cdots+\frac{1}{2}+\frac{1}{1}\right] \approx 1321$$

The sum has 221 terms and I used a computer (actually a programmable hand calculator) to evaluate it.

There is a clever argument using calculus that will give us an excellent estimate of this sum for large numbers of terms. The result is that for large N:

$$\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{N} \approx \ln(N+1)+\frac{1}{2}$$

and in fact the sum on the left is a tiny bit bigger than the expression on the right. The estimate this gives us for the given expression is:

$$221\left[\ln(222)+\frac{1}{2}\right] = 1304.5$$

That's not a bad estimate.

Problems

1. Often in games where you have to collect all the possible stickers to fill a card, they're not all equally likely to appear, in fact, typically there's one sticker that's very rare, and the goal is essentially to get that sticker. There's a version of the dice problem which captures this difficulty. Suppose the outcomes on the die are not all equally likely, so that the 6 comes up with probability 1/36, and the other five numbers come up with equal probability of 7/35 each. What now is the average number of rolls required to get every outcome at least once? Well, I'm not entirely sure how to best approach this problem--I think might be hard. But maybe a way to start thinking about it is to construct some simpler problems.

2. A jar begins with one amoeba. Every minute, every amoeba turns into 0, 1, 2, or 3 amoebae with probability 25% for each case (dies, does nothing, splits into 2, or splits into 3). What is the probability that the amoeba population eventually dies out?
[There's an important observation about "independence" to be made here. The reproductive chances of any amoeba are independent of what the other amoeba are doing. What this means is that if p is the probability that a population of exactly one amoeba will die out, then the probability that a population of exactly two amoebae will die out is p^2, and for three amoebae it's p^3, etc. Set up a tree.

3. At a movie theater, the manager announces that they will give a free ticket to the first person in line whose birthday is the same as someone who has already bought a ticket. You have the option of getting in line at any time. Assuming that you don't know anyone else's birthday, that birthdays are distributed randomly throughout the year, etc., what position in line gives you the greatest chance of being the first duplicate birthday?
[Try a dice analogue of this first (essentially assuming that the year has six days). Let p_n be the probability that the first n rolls of the dice will all be different and the $n+1^{st}$ will be a repeat of one of the first n. Write a formula for each p_n for n=1 to 6. As a check, verify with your calculator that the sum of these six numbers is 1—as it has to be. Now solve the problem for the 6-day year. Now to extend your answer to the 365-day year, you'll have to look carefully at how each p_n is obtained from the one before.