

Sons and Daughters

Martin Oldfield

30 Mar 2011 (updated).

Abstract

One of the Internet's recurring memes is a world where families have children until they have a son, then stop. There seems to be a fairly common misapprehension that this will lead to some asymmetry of sexes in the next generation.

This short note explores this question: we quickly see that symmetry is preserved, but then spend time exploring what's going on in more detail. Finally we introduce a new element to the model which, on its own, favours neither boys nor girls, but allows social engineering to skew the sex distribution.

1 Introduction

As is usually the case with this sort of probability puzzle, the main problem is to find the right way of looking at the problem. Once you've found that, the answer's obvious.

In this problem we're asked to think imagine families who have children until they have a son, then stop. As posed it's natural to think about what happens on a family-by-family basis, but that's a mistake!

Rather, enlightenment comes when you think about each birth individually: although families might be indulging in social engineering, nothing changes the statistics of the birth itself. Every baby is equally likely to be a boy or a girl—we'll ignore the observed fact that actually about 51% of babies are male.

The key insight is that the next generation is just the total of all the births: if half of the births are male, then the next generation will be half-male too. Similarly even at the family level, we'd expect to see the same number of boys and girls in each family.

Moreover, we know that every family has exactly one son in it and so *on average* it must have one daughter too. However, there's no symmetry which relates the *distributions* of sons and daughters, so the number of daughters will vary.

1.1 Simulation

Given how easy it is to mislead oneself with this sort of problem, I think it's usually sensible to simulate it. Often it's easy to mechanically generate a single sample of what's happening, and the computer is good at doing this many times then averaging the results. Happily modern computers are fast enough to simulate simple situations in a fraction of a second, even when the program is written in a slow language like Perl.

I think there are three advantages to doing the simulation:

- By forcing us to explicitly model what's going on, it exposes gaps or contradictions in our assumptions.
- By mechanically calculating the consequences for one particular random choice at a time, it makes it less likely we'll make a mistake when thinking about correlations.
- In general writing the program is a different sort of thinking to doing the mathematical analysis, so the chances of making the same error are quite small.

Of course, there are potential problems too. One is that it may not be clear how many samples one needs to take. A simple check is to simply run the simulation three times and check that the outcomes are roughly the same: if they're not something's wrong. However the converse sadly doesn't apply: if we were simulating the lottery (say a-million-to-one shot) with three 100,000 sample runs there's a reasonable chance we'd conclude that nobody ever won and you shouldn't play. Perhaps that error would be a good thing!

Here, for each family we simply simulate the family by picking sexes at random—each equally likely—until we produce a son. Then, we forget about that family and do another one. All we have to do is keep track of the number of sons and daughters, and perhaps how often we see a particular shape of family.

Explicitly, the heart of the program looks something like this¹:

```
n_boys = 0;
n_girls = 0;
do {
  if (rand() < 0.5) { n_boys++; }
  else { n_girls++; }
} while (n_boys == 0);
```

I hope that by making it quite explicit that every birth has an equal chance of being a son or a daughter, that it's clear that the expected numbers of sons and daughters in the next generation are the same. In fact, once you've written this bit of code, it's debatable whether you actually need to run it!

On the other hand, it's but a single command to run it, so I simulated 100,000 random families (which took all of about 0.3s on my laptop). Table 1 shows the results after running the program three times. I think it's clear that there's no asymmetry!

Fraction of sons	50.2%	49.9%	49.9%
Average number of sons per family	1.000	1.000	1.000
Average number of daughters per family	0.994	1.004	1.003
Average number of children per family	1.994	2.004	2.003

Table 1: Three simulations, each of 100,000 families.

1.2 Family imbalances

As we said above, although the total number of sons in the next generation will be about the same as the total number of daughters, this doesn't mean that the distribution of boys per family will match the distribution of girls. For example, about half the families will have no daughters (because the first-child was a son); on the other hand *all* the families will have exactly one son.

Figure 1 shows how families grow. We assume that every family starts out with no children and represent this by the $[0, 0]$ node at the top of the diagram.

Every time a new child is born, we move down a row in the diagram. If the child's male then we move down and left along the S arrow; if the child's female we move down and right along the D arrow.

Each node is labelled with a pair of numbers $[s, d]$ which shows the numbers of sons and daughters. Given our rule that families stop growing when the first son is born, none of the $[1, d]$

¹If you don't program, `rand()` returns a random number between 0 and 1, and `n_boys++` adds one to the number of boys. The `do { ... }` while thing repeats the process until a son is born.

nodes have arrows leading from them: these nodes correspond to final states of the family. Conversely, all the $[0, d]$ states have two arrows leading from them which correspond to the birth of a new child.

Parenthetically, if this were a model of real families all the nodes would have arrows leading from them, but the total chance of taking an arrow would be less than one: the family might stop growing at any stage.

Returning to our one-son-per-family model, we can immediately see several things:

- All families have one son.
- $\frac{1}{2}$ of the families have no daughters.
- $\frac{1}{4}$ of the families have one daughter.
- $\frac{1}{4}$ of the families have more than one daughter.

More quantitatively we can use the diagram to help us calculate the chances of any particular pattern of sons and daughters. To do this, note that every time we make a choice we choose the S or D arrow with equal chances. So, to find out the chance of getting to a particular state we just count the number of arrows we need to follow to get from the start to that node. Each step has a 50% chance of being taken, so we just raise $\frac{1}{2}$ to the relevant power to get the probability.

For example, we have to follow three arrows to get from $[0,0]$ to $[1,2]$, so the chance of getting one son and two daughters is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$ i.e. $\frac{1}{8}$.

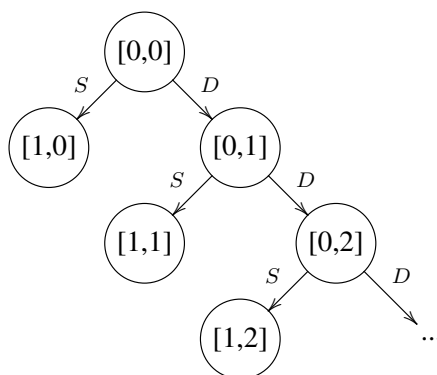


Figure 1: Possible families. Each node is labelled $[s,d]$: s is the number of sons, d the number of daughters.

Although we've already reasoned that the expected number of daughters in each family must be one, we can verify this in three different ways. These verifications aren't really meant to convince you that the answer really is one, rather they illustrate how the apparently asymmetric situation cunningly manages to be consistent with the symmetric result.

1. By recursion

Suppose d is the expected number of daughters. After the first child is born, we know there are two possibilities:

- A son was born and so we stop without any daughters.
- A daughter was born. This basically leaves us where we started but with one daughter already in the family. To see this, cover up the top node in the picture and observe that the picture hasn't changed much (except that all the d numbers have gone up by one).

Hence,

$$d = \frac{1}{2} \times 0 + \frac{1}{2} \times (d + 1), \quad (1)$$

$$d = 1. \quad (2)$$

2. By summation over final states

Consider all the nodes where a son has just been born: these are the nodes at the end of the S arrows, and correspond to the final state of the family.

If we sum over all these states then

$$d = \frac{1}{2} \times 0 + \frac{1}{4} \times 1 + \frac{1}{8} \times 2 + \dots, \quad (3)$$

$$d = \frac{1}{2} \sum_{i=0}^{\infty} i \left(\frac{1}{2}\right)^i, \quad (4)$$

$$d = 1. \quad (5)$$

The sum is a standard one you can find in tables or ask Mathematica about. However, ignoring issues like convergence, there's a cute trick to sum it:

$$\sum_{i=0}^{\infty} i\theta^i = \sum_{i=0}^{\infty} \theta \frac{d}{d\theta} \theta^i, \quad (6)$$

$$= \theta \frac{d}{d\theta} \sum_{i=0}^{\infty} \theta^i, \quad (7)$$

$$= \theta \frac{d}{d\theta} \left(\frac{1}{1-\theta} \right), \quad (8)$$

$$= \frac{\theta}{(1-\theta)^2}. \quad (9)$$

This is just like the old trick of 'differentiating-under-the-integral-sign' which Feynman talks about in 'Surely You're Joking, Mr Feynman!'.

3. By summation over births

If we consider each row of the tree, then we can ask how much it contributes to the expected number of daughters.

The probability of getting to the i^{th} row is just $(1/2)^i$ and half the time when we move down a row we'll welcome another daughter to the family. So, moving from row 0 to row 1 adds half a daughter, from row 1 to 2 adds a quarter, and so on:

$$d = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots, \quad (10)$$

$$d = 1. \quad (11)$$

One could of course calculate the expected number of sons in exactly the same way, and get the same result.

Incidentally for a quick way to do this sum, just write the number in binary: 0.111111̇.

1.3 Finite families

Figure 1 also helps us understand what happens if we limit the total number of children. The tree no longer goes on forever, but stops after a fixed number of rows. Figure 2 shows the table after these changes.

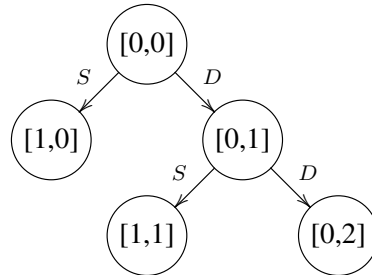


Figure 2: Families with no more than two children.

I hope it's clear that there's still no asymmetry in the expected number of sons and daughters but if you're not sure, then table 2 enumerates the three possibilities.

Final State	Family	Probability	Sons	Daughters
[1, 0]	One son	1/2	1	0
[1, 1]	One son, one daughter	1/4	1	1
[0, 2]	Two daughters	1/4	0	2
Expected value			3/4	3/4

Table 2: The three possible outcomes for families with no more than two children.

If we wanted to extend our simulation program to handle this case, it would be easy:

```

n_boys = 0;
n_girls = 0;
do {
  if (rand() < 0.5) { n_boys++; }
  else { n_girls++; }
} while (n_boys == 0 && (n_boys + n_girls <= 2));
  
```

Again it's easy to see that the central assumption—on average we produce as many sons as daughters—hasn't changed.

2 Breaking the symmetry

Having shown that the basic model won't produce a sex-asymmetry, it's natural to ask what will. Obviously we could just change the proportion of sons born, but that's rather crude and not really in the spirit of the original model. It would be nice if we could find a model which by itself produces equal numbers of sons and daughters, but which generates an asymmetry when coupled with the 'one-son-per-family' rule.

Happily we can do this! Simply suppose that some families are predisposed to have daughters and others sons—I should say that this idea is wholly hypothetical and any agreement with real biology is entirely accidental.

In particular assume that half the families have a probability $1/2 + \Delta$ of having sons, while the other half have probability $1/2 - \Delta$. In other words Δ just sets the scale of the effect: $\Delta = 0$ makes the effect vanish, $\Delta = 1/2$ makes families either have *all* sons, or *all* daughters. However, between these extremes, the two biases will cancel each other out, and we'll be back to the 50 : 50 split in the next generation.

It's important to see that the effects will only balance if size of the boy-biased families is the same as the size of the girl-biased families. This requirement is broken when social engineering rears its ugly head. Since families stop growing as soon as they have a son, families which are biased in favour of having sons will typically be smaller than those biased in favour of daughters. So the next generation will have more girl-biased babies in it, which implies that overall the fraction of sons will fall below 50%.

2.1 Simulation

It's straightforward to extend our simulation to handle this model—and one of the advantages of doing the simulation is that we this change is so small that we can be fairly confident of doing it correctly.

```
x = (rand() < 0.5) ? 0.5 + delta : 0.5 - delta;
n_boys = 0;
n_girls = 0;
do {
    if (rand() < x) { n_boys++; }
    else { n_girls++; }
} while (n_boys == 0);
```

Table 3 shows some typical results from these simulations. We consider three different cases which correspond to different values of Δ . Happily when $\Delta = 0$ we recover the results from the last section, and as Δ grows the fraction of males in the next generation falls.

2.2 The analytic result

Having worked out what the 'right' answer is, it would be nice to find an analytic result.

For a moment, suppose the bias were fixed and calculate how many daughters we'd expect. Formally suppose the probability of getting a daughter is ϕ . Then the chance of getting exactly i daughters (and one son) is just,

$$p(d = i, s = 1) = \phi^i(1 - \phi), \quad (12)$$

and so the expected number of daughters $\langle d|\phi \rangle^2$,

$$\langle d|\phi \rangle = \sum_{i=0}^{\infty} i \times \mathbf{p}(d = i, s = 1), \quad (13)$$

$$= \sum_{i=0}^{\infty} i \phi^i (1 - \phi), \quad (14)$$

$$= (1 - \phi) \frac{\phi}{(1 - \phi)^2}, \quad (15)$$

$$= \frac{\phi}{1 - \phi}. \quad (16)$$

Having solved for a particular bias, now we have to average over the two possible biases:

$$\langle d \rangle = \frac{1}{2} (\langle d|\phi = (1/2 + \Delta) \rangle + \langle d|\phi = (1/2 - \Delta) \rangle), \quad (17)$$

$$= \frac{1}{2} \left(\frac{1/2 + \Delta}{1/2 - \Delta} + \frac{1/2 - \Delta}{1/2 + \Delta} \right), \quad (18)$$

$$= \frac{1 + 4\Delta^2}{1 - 4\Delta^2}. \quad (19)$$

Or, in terms of the fraction of boys β ,

$$\beta = \frac{1}{2} (1 - 4\Delta^2). \quad (20)$$

The final row of table 3 shows the value of this: happily it agrees with our simulations.

You'll notice that this value, like some others above diverges as $\Delta \rightarrow (1/2)$. As we discussed above this just corresponds to the case where families have either *all* sons or *all* daughters. Families in the latter class will take quite a while to have a son!

Δ		0	1/8	1/4
$\mathbf{p}(\text{son})$		{1/2, 1/2}	{3/8, 5/8}	{1/4, 3/4}
Fraction of boys in simulation	Run 1	0.500	0.469	0.377
	Run 2	0.497	0.469	0.376
	Run 3	0.498	0.470	0.375
	Average	0.499	0.469	0.376
Analytic result		1/2 = 0.5	15/16 \approx 0.4688	3/8 = 0.375

Table 3: The fraction of boys as a function of Δ . The top part of the table shows results from simulating 100,000 families numerically; the last row shows the analytic result.

2.3 Other effects

We claimed before that without the one-son-per-family rule, there wouldn't be any overall effect of this model. That's true if you look at the fraction of sons in the next generation, but there is an effect on the distribution of sons and daughters in a particular family.

² $\langle d|\phi \rangle$ is just the expected value of d given a particular value for ϕ .

Let's calculate the probabilities of different families, given a particular Δ . In every case, we'll need to average over two cases:

$$p(S) \equiv p(\text{son}) = \left\{ \left(\frac{1}{2} - \Delta \right), \left(\frac{1}{2} + \Delta \right) \right\}, \quad (21)$$

$$p(D) \equiv p(\text{daughter}) = \left\{ \left(\frac{1}{2} + \Delta \right), \left(\frac{1}{2} - \Delta \right) \right\} \quad (22)$$

Consider first families with only child: there are only two possibilities, a son or a daughter. Quantitatively:

$$p(S) = \frac{1}{2} \left(\left(\frac{1}{2} - \Delta \right) + \left(\frac{1}{2} + \Delta \right) \right), \quad (23)$$

$$= \frac{1}{2}, \quad (24)$$

$$p(D) = \frac{1}{2} \left(\left(\frac{1}{2} + \Delta \right) + \left(\frac{1}{2} - \Delta \right) \right), \quad (25)$$

$$= \frac{1}{2}. \quad (26)$$

As expected perhaps, there's nothing to see there. However, now consider two children. This time we have four cases:

$$p(SS) = \frac{1}{2} \left(\left(\frac{1}{2} - \Delta \right)^2 + \left(\frac{1}{2} + \Delta \right)^2 \right), \quad (27)$$

$$= \frac{1}{4} + \Delta^2, \quad (28)$$

$$p(SD) = \frac{1}{2} \left(\left(\frac{1}{2} - \Delta \right) \left(\frac{1}{2} + \Delta \right) + \left(\frac{1}{2} + \Delta \right) \left(\frac{1}{2} - \Delta \right) \right), \quad (29)$$

$$= \frac{1}{4} - \Delta^2, \quad (30)$$

$$p(DS) = \frac{1}{2} \left(\left(\frac{1}{2} + \Delta \right) \left(\frac{1}{2} - \Delta \right) + \left(\frac{1}{2} - \Delta \right) \left(\frac{1}{2} + \Delta \right) \right), \quad (31)$$

$$= \frac{1}{4} - \Delta^2, \quad (32)$$

$$p(DD) = \frac{1}{2} \left(\left(\frac{1}{2} + \Delta \right)^2 + \left(\frac{1}{2} - \Delta \right)^2 \right), \quad (33)$$

$$= \frac{1}{4} + \Delta^2. \quad (34)$$

Here there is an effect! Not in the difference between sons and daughters though: observe that swapping S and D in the equations above doesn't change the probability. Rather in this model as Δ increases single-sex families become more probable—this really isn't a surprise.

Although we won't do it here, if you considered bigger families then the effect would become more pronounced.

In principle these results provide a signature which you could test empirically. However, to match these results against real-world data you'd need to model the non-uniform distribution of sexes, and think about twins and other multiple-births. For example, about one-in-a-thousand deliveries are identical twins, and they're almost always of the same sex.

One would either have to control for that, or perhaps regard twins as a special case of the general phenomenon this model captures. It's not quite the same though, because it's hard to have only one child when twins are born!

3 Conclusions

I hope it's clear that choosing to stop having children as soon as you've had a son won't affect the fraction of men in the next generation. Indeed I hope it's clear that no choice which only affects how many children people have will change the sex ratio assuming that we can treat each birth as an independent random event.

However, logically that assumption might not be true. We considered a simple way in which nature might be different, and saw that it couples the choice of when to stop having children to the sex of the children produced.

Although there's nothing very deep in any of this, I still think it's a nice problem.

Finally, you can download [the program I used to run the simulations](#).