What is the expected proportion of girls?

Thomas Bayes

> *There's a certain country where everybody wants to have a son. Therefore each couple keeps having children until they have a boy; then they stop. What fraction of the population is female?*
>
> *Well, of course, you can't know for sure, because, by some extraordinary coincidence, the last 100,000 families in a row might have gotten boys on the first try. But in* ***expectation****, what fraction of the population is female? In other words, if there were many such countries, what fraction would you expect to observe on average?*

This is the question – exactly as stated – that Steve Landsburg posed on his *The Big Questions* blog[1]. The subsequent discussion generated significant controversy, confusion, and conflict. There are several ways to complicate this question with vague and cloudy notions, and the comments revealed many of them. But a simple model for the births within each generation of families can reveal much of the reasoning that caused the confusion, and *hopefully* help people resolve their logical conflicts.

We can gain considerable insight into this puzzle by considering gender statistics for the children born to each generation within the country. For the first generation of children, we can envision a sequence of numbers like this:

$$G_{1,1} \quad G_{1,2} \quad G_{1,3} \quad G_{1,4} \quad \cdots \tag{1}$$

where $G_{1,k}$ is equal to 1 if the kth child of the first generation is a girl, and $G_{1,k}$ is equal to 0 if the kth child of the generation is a boy. For the second generation, we would have another sequence like this:

$$G_{2,1} \quad G_{2,2} \quad G_{2,3} \quad G_{2,4} \quad \cdots \tag{2}$$

Subsequent generations would generate similar sequences.

For any country with any birth policy, we should assume that the next child born in the generation is equally likely to be a boy or a girl regardless of the past history of boys and girls. This is a biological constraint that is independent of any policy the families in a country might use to decide how many children they will have.

[1] http://www.thebigquestions.com/2010/12/21/are-you-smarter-than-google/

Expected proportion of females for a single generation

Let's begin by thinking about a single generation of children. If the first child born to the generation is a girl, the second a boy, the third a boy, and the fourth a girl, then the gender sequence for that generation will begin like this:

$$1 \quad 0 \quad 0 \quad 1 \quad \cdots, \tag{3}$$

and the gender sequence for the first ten children might look like this:

$$1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad \cdots \tag{4}$$

Eventually the last child for the generation will be born, and, if the country has no particular birth policy, then the sequence could terminate with either a boy or a girl. A complete sequence might look like this:

$$1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \tag{5}$$

but, of course, most countries will have many more than 13 children in a particular generation.

A country with no birth policy

For a country with no particular birth policy, every child born to a generation (the first, the second, the third, . . ., and the last) is equally likely to be a boy or girl. In other words, the sequence of births will be statistically equivalent to a sequence we might obtain by repeatedly tossing a fair coin. Because of this, the expected proportion of females for any total number of children will be equal to 1/2. This will be true if we look at the first child born to the generation, the first 10 children, or all of the generation's children. We can easily verify this result, and the value of 1/2 matches well with most people's intuition.

A systematic census error

Beginning with a comment in response to the *Win Landsburg's Money!!!* post[2], commenter Tom pointed out that each family in a country where every family wants to have a boy will terminate its sequence of births with a boy, and, as a result, the sequence of births for each generation of families will also terminate with a boy[3]. To fully appreciate the ultimate significance of this characteristic of the country's birth policy, we start with a simplified model and work our way toward the important final result.

[2] http://www.thebigquestions.com/2010/12/27/win-landsburgs-money/
[3] http://www.thebigquestions.com/2010/12/27/win-landsburgs-money/#comment-19691

Suppose that anytime we take a census for a country we always record the youngest child in our count as a boy. If we count only the first 4 children, for example, then an actual birth sequence that looked like this

$$1 \quad 0 \quad 0 \quad 1, \tag{6}$$

would be recorded as this

$$1 \quad 0 \quad 0 \quad 0, \tag{7}$$

but a sequence that looked like this

$$1 \quad 0 \quad 1 \quad 0, \tag{8}$$

would be recorded without error. What is the expected proportion of females for our census?

If we count K children in our census, then the expected number of females in our count will be $(K-1)/2$, because the gender for the first $K-1$ children will be recorded accurately, and the gender for the Kth child will be recorded as a male. The expected number of males in our count will be $(K-1)/2+1$ which is equal to $K/2+1/2$, so, as suggested by commenter Tom[4], we might refer to this type of error as introducing an *extra half-boy* to our census count. Because of this *extra half-boy*, the expected proportion of females in the census will be

$$E[\text{proportion of females}] = \frac{(K-1)/2}{K} = \frac{1}{2} - \frac{1}{2K}, \tag{9}$$

which is an exact expression for the expected proportion of females in the presence of this systematic counting error, and, although the expected proportion will be close to $1/2$ for very large K, it would be wrong to declare that it is *exactly* equal to $1/2$. It is not.

A country where the last child born in a generation is replaced with a boy

Suppose a country has no particular birth policy, but the last child born to the generation is miraculously replaced with a boy. The expected proportion of females in this country will now depend on the number of children we include in the census. If we compute the proportion of females before the last child is born, then the expected proportion will be exactly $1/2$, and, if we compute the proportion of females after the last child is born, then the expected proportion will be $1/2 - 1/(2K)$ because we will have to account for the *extra half-boy*. But the total number of children K that the generation will have is likely to be random, so the expected proportion will be more precisely $1/2 - E[1/(2K)]$, where we've replaced $1/(2K)$ with its expected value. This expected value will, in turn, depend primarily on two factors: the number of families that make up the generation; and the distribution for the number of children born to each family.

[4] http://www.thebigquestions.com/2010/12/27/win-landsburgs-money/#comment-19736

If we don't know whether or not all of the generation's children will have been born at the time of our census, then the expected proportion of females within this country will be

$$\begin{aligned} E[\text{proportion of females}] &= \frac{1}{2}\Pr[\text{not complete}] + \left(\frac{1}{2} - E\left[\frac{1}{2K}\right]\right)\Pr[\text{complete}] \\ &= \frac{1}{2} - E\left[\frac{1}{2K}\right]\Pr[\text{complete}], \end{aligned} \tag{10}$$

where Pr[complete] is the probability that we will take the census after the generation has completed all of its births. Unless we can be certain that all families have not completed their births, then the expected proportion of females cannot be *exactly* 1/2.

A country where every family wants a boy

In a country where every family wants a boy, the gender sequence for a particular generation will always end with a boy. Unlike the previous example, however, the last boy will come about from a deliberate policy that all of the families have adopted. But, other than that, the analysis we used to understand the previous situation can apply here. Furthermore, because we know more about the birth policy, we can provide a little more insight into the expected value for $1/(2K)$.

The total number of boys born to a generation will be equal to the number of families, and the total number of girls born to the generation will be a random variable with a negative binomial distribution whose expected value is equal to the number of families, and whose variance is equal to twice the number of families[5]. The expected number of children, then, will be equal to twice the number of families, and the variance of the number of children will also be equal to twice the number of families. To determine the expected value of $1/(2K)$ we can use a Taylor series expansion about the mean of K:

$$\frac{1}{2K} \simeq \frac{1}{4N} - \frac{1}{8N^2}(K-2N) + \frac{1}{16N^3}(K-2N)^2, \tag{11}$$

so that

$$\begin{aligned} E\left[\frac{1}{2K}\right] &\simeq \frac{1}{4N} - \frac{1}{8N^2}E[(K-2N)] + \frac{1}{16N^3}E\left[(K-2N)^2\right] \\ &= \frac{1}{4N} + \frac{1}{16N^3}2N \\ &= \frac{1}{4N} + \frac{1}{8N^2} \\ &\simeq \frac{1}{4N}, \end{aligned} \tag{12}$$

where N is the number of families in the generation. Of course if we don't know N, then we will

[5] http://en.wikipedia.org/wiki/Negative_binomial_distribution

need to replace $1/(4N)$ with its expected value.

So the expected proportion of females in a country where every family wants a boy is roughly equal to

$$E[\text{proportion of females}] = \frac{1}{2} - E\left[\frac{1}{4N}\right] \Pr[\text{complete}], \tag{13}$$

where N is equal to the number of families in the generation, and Pr[complete] is the probability that all of the families have had all of their children at the time we take the census. The only way to make this expected value *exactly* equal to $1/2$ is to ensure that at least one family has not had a boy at the time of the census. But the original question is about the expected value of the proportion of females, so there seems to be no way to be certain that all families have not had their boy at the time of the census. There is a nonzero probability, after all, that all of the families will have their boy on the first try. So, although it might be small, it is unlikely that Pr[complete] will be zero, so the expected proportion of females will not be *exactly* $1/2$ because there will be some chance that our census will include the *extra half-boy*.

As part of the discussion about this puzzle, commenter Tom suggested that we avoid this issue by omitting the youngest child from our census[6]. If our census includes 10 children, we only count the 9 oldest. If our census includes one million children, we only count the $999,999$ oldest. By doing this, our census will never include a complete generation of children, and Pr[complete] will be zero. Whereas this biased census will ensure that the *extra half-boy* is never included in the count, we will not have a precise measure of the proportion of females in the country. Any unbiased census we retain a nonzero chance of including the *extra half-boy* in its count, and this is precisely why the expected proportion of females is not exactly equal to $1/2$.

Multiple generations

At this point, a complete answer to the question could address the issue of including multiple generations in the census. Accounting for multiple generations, the proportion of females in a country would be

$$\text{proportion of females} = \frac{G_1 + G_2 + G_3 + \cdots}{C_1 + C_2 + C_3 + \cdots}, \tag{14}$$

where G_m is the number of girls counted from the mth generation, and C_m is the number of children counted from the mth generation. We know that the expected ratio of G_m/C_m is smaller than $1/2$ for any particular generation, but the question would remain regarding the expected ratio of the sums. I haven't addressed this issue, but I believe it is unlikely that something magic will happen to make the expected value of this ratio *exactly* equal to $1/2$. Anyone care to tighten this up?

[6] http://www.thebigquestions.com/2010/12/27/win-landsburgs-money/#comment-20076