

I have two children, not both girls. Do I have a daughter?

C. M. Caves
2009 January 18

1. The original problem

In the original statement of the problem, I give you the initial information I : *I have two children*. This information sets up the relevant hypothesis space: the space of two children, each of whom can be either a boy (b) or a girl (g). Since *all* probabilities in the problem are conditioned on I , we can omit that conditioning without losing anything. We assume that you start with prior probabilities, $P(b) = b$ and $P(g) = g$, for each child to be a boy or girl. Usually, you would assume that $b = g = 1/2$, but we allow the more general situation just to see how a prejudice toward boys or girls enters into the final results. We also assume that you consider the two births to be independent, so that the four birth sequences have probabilities $P(bb) = b^2$, $P(bg) = bg$, $P(gb) = gb$, and $P(gg) = g^2$, where the left slot is the older child.

You are now given the information I_1 : *They're not both girls*. You are asked for the probability that I have a daughter. The conditional probabilities for I_1 , given the birth sequences, are

$$P(I_1|bb) = P(I_1|bg) = P(I_1|gb) = 1, \quad P(I_1|gg) = 0, \quad (1)$$

with

$$P(I_1) = \sum_x P(I_1|x)P(x) = b^2 + 2bg, \quad (2)$$

where x denotes the birth sequences. Bayes's theorem, as one would expect, simply rules out gg and renormalizes the other three birth sequences:

$$P(x|I_1) = \frac{P(I_1|x)P(x)}{P(I_1)} = \begin{cases} b^2/(b^2 + 2bg), & x = bb, \\ bg/(b^2 + 2bg), & x = bg \text{ or } gb, \\ 0, & x = gg. \end{cases} \quad (3)$$

The probability for a daughter is

$$P(d|I_1) = P(bg|I_1) + P(gb|I_1) = \frac{2bg}{b^2 + 2bg} = \frac{2g}{1 + g}. \quad (4)$$

When $b = g = 1/2$, this gives $2/3$ probability for a daughter, the key point being that there are two birth sequences that give a daughter and only one that doesn't.

2. Additional information

Now suppose I give you additional information I_2 . You account for it by introducing the conditional probabilities for I_2 , $P(I_2|x)$, given the birth sequences x . It might seem that these probabilities should also be conditioned on I_1 , but it is hard to imagine a situation where this is necessary, since any dependence on whether or not there are two

girls is already incorporated in the conditioning on birth sequences. Bayes's theorem now gives

$$\begin{aligned}
P(x|I_2, I_1) &= \frac{P(I_2|x)P(x|I_1)}{P(I_2|I_1)} \\
&= \frac{P(I_2|x)P(I_1|x)P(x)}{P(I_2, I_1)} \\
&= \frac{1}{P(I_2|bb)b^2 + [P(I_2|bg) + P(I_2|gb)]bg} \times \begin{cases} P(I_2|bb)b^2, & x = bb, \\ P(I_2|bg)bg, & x = bg, \\ P(I_2|gb)bg, & x = gb, \\ 0, & x = gg, \end{cases}
\end{aligned} \tag{5}$$

where

$$P(I_2|I_1) = \sum_x P(I_2|x)P(x|I_1) = \frac{P(I_2|bb)b^2 + [P(I_2|bg) + P(I_2|gb)]bg}{b^2 + 2bg}. \tag{6}$$

The second form in Eq. (5) shows that you can apply I_1 and I_2 in either order, and the final form displays clearly that the entire effect of I_1 is to rule out the birth sequence gg , thus rendering the conditional probability $P(I_2|gg)$ irrelevant.

After receiving the information I_2 , you report the probability for a daughter to be

$$\begin{aligned}
P(d|I_2, I_1) &= P(bg|I_2, I_1) + P(gb|I_2, I_1) = \frac{2Qbg}{Pb^2 + 2Qbg} \\
&= \frac{2Qg}{Pb + 2Qg} \\
&= \frac{Q}{P} \frac{2g}{1 + (2Q/P - 1)g},
\end{aligned} \tag{7}$$

where

$$P \equiv P(I_2|bb), \quad Q \equiv \frac{1}{2}[P(I_2|bg) + P(I_2|gb)]. \tag{8}$$

The original problem corresponds to I_2 being vacuous, i.e., $P(I_2|x) = 1$ for all four values of x . If $P = 2Q$, you have $P(d|I_2, I_1) = g$.

The most general statement we can make about whether I_2 has any effect is that $P(d|I_2, I_1) = P(d|I_1)$ if and only if $P = Q$. The probability for a daughter changes when I_2 breaks this equality; we call such information *cogent*. We note immediately that P and Q are invariant under exchange of birth order, which implies that any birth-order preference is not cogent.

There are several kinds of symmetries that are relevant for the conditional probabilities.

1. I_2 makes no gender distinction on the younger child: the conditional probabilities $P(I_2|x)$ are invariant under a gender flip in the younger (right) slot, i.e.,

$$P(I_2|bb) = P(I_2|bg), \quad P(I_2|gb) = P(I_2|gg). \tag{9}$$

If $P(I_2|bb) = P(I_2|bg) \neq P(I_2|gb) = P(I_2|gg)$, then $P \neq Q$, and I_2 provides cogent information about whether I have a daughter. The cogent information can be thought as arising from a gender distinction for the older child.

2. I_2 makes no gender distinction on the older child: the conditional probabilities $P(I_2|x)$ are invariant under a gender flip in the older (left) slot, i.e.,

$$P(I_2|bb) = P(I_2|gb), \quad P(I_2|bg) = P(I_2|gg). \quad (10)$$

If $P(I_2|bb) = P(I_2|gb) \neq P(I_2|bg) = P(I_2|gg)$, then $P \neq Q$, and I_2 provides cogent information about whether I have a daughter. The cogent information can be thought as arising from a gender distinction for the younger child.

3. I_2 makes no gender distinction: the conditional probabilities $P(I_2|x)$ are invariant under all gender flips and thus are independent of x . Not surprisingly, I_2 has no effect on the probability I have a daughter.
4. I_2 makes no birth-order distinction: the conditional probabilities $P(I_2|x)$ are invariant under exchange of birth order, i.e.,

$$P(I_2|bg) = P(I_2|gb). \quad (11)$$

If $P(I_2|bb) \neq P(I_2|bg) = P(I_2|gb)$, then $P \neq Q$, and I_2 provides cogent information about whether I have a daughter. The cogent information comes from a gender distinction that is birth-order invariant. This is the most interesting sort of symmetry. The information I_2 is cogent when you believe that the birth sequence bb makes I_2 more or less likely than for the birth sequences bg and gb.

The gender distinction introduced by I_2 is quantified by $P - Q$. It is convenient to introduce

$$\delta \equiv P(I_2|bg) - P(I_2|gb). \quad (12)$$

The quantity

$$\begin{aligned} \Delta &\equiv \frac{|P - P(I_2|gb)| - |P - P(I_2|bg)|}{2|P - Q|} \\ &= \begin{cases} \delta/2|P - Q|, & P \geq \max[P(I_2|bg), P(I_2|gb)], \\ -\delta/2|P - Q|, & P \leq \min[P(I_2|bg), P(I_2|gb)], \\ 2(P - Q)/|P - Q|, & P(I_2|gb) \leq P \leq P(I_2|bg), \\ 2(Q - P)/|P - Q|, & P(I_2|bg) \leq P \leq P(I_2|gb). \end{cases} \end{aligned} \quad (13)$$

can be regarded as quantifying the birth-order contribution to any gender distinction made by I_2 . When $\Delta = 0$, any gender distinction is birth-order invariant. When Δ is positive, there is a weighting toward a gender distinction on the younger child, and when Δ is negative, there is a weighting toward a gender distinction on the older child.

3. Examples

Some examples help to make all this clearer. In considering the examples, we give the conditional probability $P(I_2|gg)$, even though it is irrelevant to the inferences conditioned

on I_1 . It would come into play if we applied I_2 first, but would disappear once I_1 was taken into account.

1. I_2 : *The older child is a boy.* In this example, you have $P(I_2|bb) = P(I_2|bg) = 1$ and $P(I_2|gb) = P(I_2|gg) = 0$ and thus $P(d|I_2, I_1) = g$. This is an example of symmetry 1, with the cogent information coming from a gender distinction for the older child.

2. I_2 : *They're both here in the room somewhere. Ah, there's one of them now, my son Bob.* In this example, $P(I_2|x)$ is the probability that the first child we encounter in the room is a boy, given birth sequence x , so you have $P(I_2|bb) = 1 - P(I_2|gg) = 1$ and $P(I_2|bg) = P(I_2|gb) = 1/2$ and thus $P(d|I_2, I_1) = g$. This is an example of symmetry 4.

3. I_2 : *They're both here in the room somewhere. Ah, there's one of them now, my daughter Sue.* This one is utterly trivial, because I_2 tells you directly that I have a daughter (assuming the standard naming conventions). Formally, $P(I_2|x)$ is the probability that the first child we encounter is a girl, so you have $P(I_2|bb) = 1 - P(I_2|gg) = 0$ and $P(I_2|bg) = P(I_2|gb) = 1/2$ and thus $P(d|I_2, I_1) = 1$. This is an example of symmetry 4.

4. I_2 : *The better athlete is a boy.* Let's suppose you have a prejudice that a boy is a better athlete than a girl with probability q . You would assign probabilities $P(I_2|bb) = 1 - P(I_2|gg) = 1$ and $P(I_2|bg) = P(I_2|gb) = q$ and thus get

$$P(d|I_2, I_1) = \frac{2qbg}{b^2 + 2qbg} = \frac{2qg}{1 + (2q - 1)g}. \quad (14)$$

Notice that if you think boys are always better athletes ($q = 1$), then I_2 provides no new information, because all I have done is to repeat that I have at least one son. In contrast, if you think girls are always better athletes ($q = 0$), then you think that I could not have made the statement I_2 for the sequences bg and gb , so you conclude that I do not have a daughter. If you have no prejudice about the comparative athletic abilities of boys and girls ($q = 1/2$), you report the probability $P(d|I_2, I_1) = g$. This is an example of symmetry 4.

5. I_2 : *The older child is the better athlete.* There is an additional complication in this example, because there is a stronger form of birth-order invariance. We get at by considering the conditional probabilities for the negation of I_2 , i.e., $P(-I_2|x)$, where $-I_2$ is the statement that *the younger child is the better athlete*. In this situation, making no birth-order distinction requires, in addition to symmetry 4, that $P(I_2|x) = P(-I_2|\bar{x})$, where \bar{x} denotes the swapped birth sequence. For this to be true, we must have $P(I_2|x) = 1 - P(I_2|\bar{x}) = 1 - P(I_2|x)$, implying that $P(I_2|x) = 1/2$ for all four values of x .

In this example, you might have both gender prejudices and the more general birth-order prejudices about which child is the better athlete, and these would most generally be expressed by adjusting all four conditional probabilities $P(I_2|x)$, thus making this an example of the most general inference. There is no unambiguous way to sort out whether you have birth-order or gender preferences from the values of your conditional probabilities.

Suppose, for example, that you assign conditional probabilities $P(I_2|bb) = P(I_2|gg) = r$ and $P(I_2|bg) = 1 - P(I_2|gb) = q$. We can think of r as describing a birth-order preference for same-sex children, whereas q describes some combination of birth-order and gender preferences for children of opposite sex. In this situation, you end up with probability $P(d|I_2, I_1) = bg/(rb^2 + bg) = (1 + rb/g)^{-1}$. If you think the older child of a same-sex pair is always the better athlete ($r = 1$), then you conclude that $P(d|I_2, I_1) = g$. If you think

the younger child of a same-sex pair is always the better athlete ($r = 0$), then you believe that I could not make the statement I_2 for the same-sex birth sequences, and thus you conclude that I do have a daughter.

6. I_2 : *One and only one of them—and that one’s a boy—has red hair.* Let’s suppose you assume that having red hair is uncorrelated with gender. A convenient way to parameterize the probabilities for red hair is

$$\begin{aligned} P(\neg\text{red}, \text{red}|x) &= P(\text{red}, \neg\text{red}|x) = q(1 - r) , \\ P(\text{red}, \text{red}|x) &= qr , \\ P(\neg\text{red}, \neg\text{red}|x) &= 1 - 2q + qr , \end{aligned} \tag{15}$$

with q and r satisfying $0 \leq q, r \leq 1$ and $q(2 - r) \leq 1$. We can think of q as being the probability for a single child to have red hair and r as being the conditional probability for a second child also to have red hair. When $r = q$, the two children’s hair color is uncorrelated. For $q \leq r \leq 1$, red hair is correlated, with perfect correlation for $r = 1$. For $r \leq q \leq 1/(2 - r)$, red hair is anticorrelated, with perfect anticorrelation for $r = 0$.

You now assign conditional probabilities

$$P(I_2|bb) = 2q(1 - r) , \quad P(I_2|bg) = P(I_2|gb) = q(1 - r) , \quad P(I_2|gg) = 0 , \tag{16}$$

giving $P = 2Q = 2q(1 - r)$ and thus $P(d|I_2, I_1) = g$. This example works just like example 2, making $P(I_2|bb)$ twice as big as $P(I_2|bg) = P(I_2|gb)$ simply because having two boys makes for two ways for a single boy to have red hair. The correlation of successive children is irrelevant to your inference. This is an example of symmetry 4.

7. I_2 : *One of them is a red-headed boy.* Making the same assumptions as in the preceding example, you would assign conditional probabilities

$$\begin{aligned} P(I_2|bb) &= 2q(1 - r) + qr = 2q - qr , \\ P(I_2|bg) &= P(I_2|gb) = q(1 - r) + qr = q , \\ P(I_2|gg) &= 0 , \end{aligned} \tag{17}$$

giving $P = 2q - qr$ and $Q = q$ and thus

$$P(d|I_2, I_1) = \frac{2qg}{(2q - qr)b + 2qg} = \frac{g}{1 - rb/2} . \tag{18}$$

Now everything depends on the correlation r . For perfect correlation ($r = 1$), I_2 isn’t cogent because telling you that I have one red-haired child implies I have two, so all I have done is to repeat that I have at least one boy. Any reduction in correlation makes the information cogent by increasing the probability that two boys satisfy I_2 relative to the probability for a mixed-sex pair. For perfect anticorrelation ($r = 0$), you’re back in the preceding example, because I can’t have two red-haired children. Since the probability for an initial red-haired child is actually quite small, in the absence of correlation ($r = q$), you’re nearly in the situation of perfect anticorrelation.

8. I_2 : *They run a cafe called Bob and Sue's 50s Diner.* This one is completely obvious, it being clear from the outset, given standard naming conventions (which we assume here, since the next example illustrates what happens when the naming conventions are ambiguous), that one of the children is a daughter.

To do things formally, however, let's loosen the naming conventions, in preparation for more general examples. In general, you would start with the probabilities that a particular name is a boy or a girl, i.e., the probabilities $P(b|Bob) = 1 - P(g|Bob)$ and $P(g|Sue) = 1 - P(b|Sue)$. From these, Bayes's rule gives the conditional probabilities

$$\begin{aligned}
 q = P(\text{Bob}|b) &= \frac{P(b|\text{Bob})P(\text{Bob})}{P(b)} , \\
 q' = P(\text{Bob}|g) &= \frac{P(g|\text{Bob})P(\text{Bob})}{P(g)} , \\
 r = P(\text{Sue}|g) &= \frac{P(g|\text{Sue})P(\text{Sue})}{P(g)} , \\
 r' = P(\text{Sue}|b) &= \frac{P(b|\text{Sue})P(\text{Sue})}{P(b)} .
 \end{aligned} \tag{19}$$

You would then assign conditional probabilities to my having children named Bob and Sue, in the indicated order:

$$\begin{aligned}
 P(\text{Bob, Sue}|bb) &= qr' , & P(\text{Sue, Bob}|bb) &= r'q , \\
 P(\text{Bob, Sue}|bg) &= qr , & P(\text{Sue, Bob}|bg) &= r'q' , \\
 P(\text{Bob, Sue}|gb) &= q'r' , & P(\text{Sue, Bob}|gb) &= rq , \\
 P(\text{Bob, Sue}|gg) &= q'r , & P(\text{Sue, Bob}|gg) &= rq' .
 \end{aligned} \tag{20}$$

Now you assign probabilities to I_2 conditioned on the two birth sequences for Bob and Sue and on their sexes. These probabilities are

$$\begin{aligned}
 P(I_2|\text{Bob, Sue; bb}) , & P(I_2|\text{Sue, Bob; bb}) , \\
 P(I_2|\text{Bob, Sue; bg}) , & P(I_2|\text{Sue, Bob; bg}) , \\
 P(I_2|\text{Bob, Sue; gb}) , & P(I_2|\text{Sue, Bob; gb}) , \\
 P(I_2|\text{Bob, Sue; gg}) , & P(I_2|\text{Sue, Bob; gg}) .
 \end{aligned} \tag{21}$$

These eight probabilities allow you to condition on all four birth sequence and also on how the two names are attached to the birth sequence, thus allowing you to express a prejudice for how the names are ordered in the cafe's name.

All this leads to

$$\begin{aligned}
P(I_2|bb) &= P(I_2|Bob, Sue; bb)P(Bob, Sue|bb) + P(I_2|Sue, Bob; bb)P(Sue, Bob|bb) \\
&= [P(I_2|Bob, Sue; bb) + P(I_2|Sue, Bob; bb)]qr' , \\
P(I_2|bg) &= P(I_2|Bob, Sue; bg)P(Bob, Sue|bg) + P(I_2|Sue, Bob; bg)P(Sue, Bob|bg) \\
&= P(I_2|Bob, Sue; bg)qr + P(I_2|Sue, Bob; bg)q'r' , \\
P(I_2|gb) &= P(I_2|Bob, Sue; gb)P(Bob, Sue|gb) + P(I_2|Sue, Bob; gb)P(Sue, Bob|gb) \\
&= P(I_2|Bob, Sue; gb)q'r' + P(I_2|Sue, Bob; gb)qr , \\
P(I_2|gg) &= P(I_2|Bob, Sue; gg)P(Bob, Sue|gg) + P(I_2|Sue, Bob; gg)P(Sue, Bob|gg) \\
&= [P(I_2|Bob, Sue; gg) + P(I_2|Sue, Bob; gg)]q'r .
\end{aligned} \tag{22}$$

Now notice that

$$\begin{aligned}
\frac{Pb^2}{P(Bob)P(Sue)} &= \frac{[P(I_2|Bob, Sue; bb) + P(I_2|Sue, Bob; bb)]qr'b^2}{P(Bob)P(Sue)} \\
&= [P(I_2|Bob, Sue; bb) + P(I_2|Sue, Bob; bb)]P(b|Bob)P(b|Sue) , \\
\frac{2Qbg}{P(Bob)P(Sue)} &= \frac{[P(I_2|Bob, Sue; bg) + P(I_2|Sue, Bob; gb)]qrbg}{P(Bob)P(Sue)} \\
&\quad + \frac{[P(I_2|Sue, Bob; bg) + P(I_2|Bob, Sue; gb)]q'r'bg}{P(Bob)P(Sue)} \\
&= [P(I_2|Bob, Sue; bg) + P(I_2|Sue, Bob; gb)]P(b|Bob)P(g|Sue) \\
&\quad + [P(I_2|Sue, Bob; bg) + P(I_2|Bob, Sue; gb)]P(b|Sue)P(g|Bob) .
\end{aligned} \tag{23}$$

These quantities can be used to construct the final result for the probability of a daughter. That they are invariant under simultaneous birth-order exchange and name exchange shows that any prejudice based on the birth order of the names is not cogent.

In the example at hand, you use strict naming conventions, so you would have $P(g|Bob) = P(b|Sue) = 0$, giving $P(d|I_2, I_1) = 1$, as expected. There is clearly a problem in the case of strict naming conventions if you insist that a girl's name come first. The source of the problem is that according to you, I simply can't have made the statement I_2 under these circumstances.

9. I_2 : *They run a cafe called Bob and Dana's 50s Diner.* In this example, given the gender ambiguity of Dana, you would use the formulation of the previous example, changing Sue to Dana and setting $P(g|Bob) = 1 - P(b|Bob) = 0$. This gives

$$P(d|I_2, I_1) = \frac{P(g|Dana)}{P(g|Dana) + P(b|Dana) \left(\frac{P(I_2|Bob, Dana; bb) + P(I_2|Dana, Bob; bb)}{P(I_2|Bob, Dana; bg) + P(I_2|Dana, Bob; gb)} \right)} . \tag{24}$$

If you believe that only two boys can establish and run a cafe, then the factor in big parentheses will go to infinity; you will be confident that Dana is a boy and thus that I have no daughter. If you believe that a boy and a girl are far more likely to establish and

run such an enterprise than two boys, then the factor in big parentheses will go to zero; you will be confident Dana is a girl and thus that I have a daughter. If you are indifferent to the genders of the two owners, then the factor in big parentheses will be equal to unity; the probability I have a daughter then reduces to the probability that Dana is a girl.

10. I_2 : *They run a cafe called Dana and Sue's 50s Diner.* This is trivial, of course, because you know immediately that I have a daughter Sue, given the standard naming conventions. Formally, you would use the formulation of example 8, changing Bob to Dana and setting $P(b|Sue) = 1 - P(g|Sue) = 0$, which gives $P(d|I_2, I_1) = 1$. What you learn from I_1 in this case is that Dana is definitely a boy.