The Two-Child Paradox Reborn?

Stephen Marks and Gary Smith

For at least half a century, probability devotees have puzzled over a problem known as the two-child paradox. Suppose we know that a family has two children, and we learn that one of them is a girl. What is the probability that there are two girls? Some have argued that the probability is one-half, others that it is one-third.

This problem has reappeared in various guises over the years. Does it matter, for example, if we learn not only the gender of one of the children, but whether the child is the older or younger sibling?

A new variant of the two-child problem was introduced recently by Leonard Mlodinow in *The Drunkard's Walk: How Randomness Rules Our Lives*. Mlodinow argues that the two-child probabilities depend on whether the child we learn about has an unusual name: The probability of two girls is one-third if we learn that one of the children is a girl, but it is one-half if we learn that her name is Florida.

This claim is paradoxical in that it does not seem that knowing a child's name should have any effect on the probability that her sibling is male or female.

We will first analyze the classic two-child problem, in which it is known only that one of the children is a girl. We will then consider the case in which it is also known that the girl is named Florida. For both of these scenarios, we will work through two approaches: a sample space approach and Bayesian analysis.

The Classic Two-Child Problem

We employ the traditional assumptions that boys and girls are equally likely and that the sexes of the children in a family are independent. To be clear about which child is which, we let BG indicate that, in a two-child family, the older child is a boy and the younger a girl, and similarly define GB, GG, and BB. Under the traditional assumptions, the probability of each of these outcomes is equal to one-fourth.

The Sample Space Approach

The conclusion that the probability of two girls is one-third follows from a restricted sample space analysis. It is based on the simple inference that, if one of the children is a girl, then there cannot be two boys. Because the three other possibilities—GB, BG, and GG—are presumed to be equally likely, the probability of GG must be one-third.

This analysis is correct if we are solving a different, hypothetical problem in which we draw a family at random from all families that are GG, BG, or GB. In this case, there is exactly a one-third probability of choosing a GG family. This conclusion is not paradoxical. However, the situation is different if we are looking at a family drawn at random from the population of all two-child families, and the difference is best revealed by Bayesian analysis.

Table 1—Bayesian Analysis: Mother Mentions Either Child Equally Often

	BB	BG	GB	GG	Total
Mentions Boy	100	50	50	0	200
Mentions Girl	0	50	50	100	200
Total	100	100	100	100	400

Table 2—Bayesian Analysis: Boys Are Mentioned Only If There Are No Girls

	BB	BG	GB	GG	Total
Mentions Boy	100	0	0	0	100
Mentions Girl	0	100	100	100	300
Total	100	100	100	100	400

Table 3—Bayesian Analysis: Mother Mentions Child Independent of Gender and Name

	BB	BG	GB	GG	Total
Mentions Girl Named Florida	0	$\alpha 50$	$\alpha 50$	$\alpha 100$	$\alpha 200$
Mentions Girl Not Named Florida	0	$(1-\alpha)50$	$(1-\alpha)50$	$(1-\alpha)100$	$(1-\alpha)200$
Mentions Boy	100	50	50	0	200
Total	100	100	100	100	400

The Bayesian Approach

The Bayesian approach begins with "prior probabilities" of one-fourth for each of the four possibilities: P(BG) = P(GB) = P(GG) = P(BB) = 1/4. It then uses Bayes' Theorem to calculate "posterior probabilities" that are conditional on the new information about the family.

One attractive feature of the Bayesian approach is that it encourages us to think about how the information was obtained. Suppose we learn the sex of one of the children because the mother mentions him or her. In this case, the mother mentions a girl, G_m, and we wish to find the conditional probability that the family is GG, given this information, $P(GG \mid G_m)$.

Applying Bayes' Theorem, which uses the multiplication and addition rules of probability, we get

$$P(GG \mid G_m) = \frac{P(GG \text{ and } G_m)}{P(G_m)}$$

$$= \frac{P(GG \text{ and } G_m)}{P(GG \text{ and } G_m) + P(BG \text{ and } G_m) + P(GB \text{ and } G_m) + P(BB \text{ and } G_m)}$$

$$= \frac{P(G_m \mid GG)\cdot P(GG)}{P(G_m \mid GG)\cdot P(GG) + P(G_m \mid BG)\cdot P(BG) + P(G_m \mid GB)\cdot P(GB) + P(G_m \mid BB)\cdot P(BB)}.$$

If the mother has two daughters, she can only mention a girl: $P(G_m \mid GG) = 1$. If she has two sons, she can only mention a boy: $P(G_m \mid BB) = 0$. If she has a son and a daughter, a plausible assumption is that she is equally likely to mention either child: $P(G_m \mid BG) = P(G_m \mid GB) = 1/2$. Plugging these conditional probabilities, and the prior probabilities, into the formula, we get $P(GG \mid G_m) = 1/2$.

The calculations can be done succinctly using a contingency table. For the case just examined, Table 1 shows the expected number of families for various contingencies, based on a random sample of 400 families with two children. The entries in the interior of the table equal joint probabilities like $P(GG \text{ and } G_m)$ multiplied by 400, while the totals equal marginal probabilities like $P(G_m)$ (for one of the rows) or P(GG) (for one of the columns) multiplied by 400.

For example, we expect there to be 100 BG families, and that the mother would mention the boy in half of these and the girl in the other half. Notice that the outcomes BG, GB, and GG are not equally likely once the mother mentions the gender of one of the children, contrary to the assumption of the sample space approach.

For example, the joint probability $P(GG \text{ and } G_m)$ equals $P(G_m \mid GG)\,P(GG) = (1)(1/4) = 1/4$. This implies an expected number of families equal to 100. It is twice the joint probability of the family being BG and the girl being mentioned, for which the expected number of families is 50. However, no matter whether the mother mentions a boy or a girl, her other child is equally likely to be a boy or a girl.

For example, if she mentions a girl, we can use the entries from the relevant row of the table to see that the probability that the other child is a boy is given by (50 + 50)/200 = 1/2, while the probability that the other child is a girl is given by 100/200 = 1/2.

One-third would be the correct answer only under the assumption that the mother never mentions a son if she has a daughter: $P(G_m \mid BG) = P(G_m \mid GB) = 1$. Based on the expected numbers of families shown in Table 2, which applies this assumption, the probability of a two-girl family equals one-third if the mother mentions a girl and the probability of a two-boy family equals one if she mentions a boy.

This extreme assumption is never included in the presentation of the two-child problem, however, and is surely not what people have in mind when they present it.

A Girl Named Florida

Mlodinow examines the two-child probabilities in the case in which it is learned one of the children is a girl named Florida, evidently an unusual name. He begins with this statement of the classic two-child problem:

> Suppose a mother is carrying fraternal twins and wants to know the odds of having two girls, a boy and a girl, and so on … In the two-daughter problem, an additional question is usually asked: What are the chances, *given that one of the children is a girl*, that both children will be girls?

His answer to the first question is the usual equal probabilities for BB, BG, GB, and GG. His answer to the second question is one-third, based on the sample-space analysis of the classic two-child problem.

As shown above, this is fine if Mlodinow does not mean that one of the two children is observed or mentioned to be a girl, but rather is answering a hypothetical question about the chances of there being two girls in families that do not have two boys.

Mlodinow later writes the following:

> [I]f you learn that one of the children is a girl named Florida … the answer is not 1 in 3—as it was in the two-daughter problem—but 1 in 2. The added information—your knowledge of the girl's name—makes a difference.

Mlodinow is not talking about hypothetical families that do not have two boys. The fraternal twins have been born, and we learn that one of the children is a girl named Florida.

Mlodinow supposes that one in a million girls are named Florida. We will generalize this slightly and suppose that a fraction α of girls is named Florida. Like Mlodinow, we suppose children are named independently, so that, in principle, a family could have two girls named Florida.

Bayesian Analysis

Table 3 shows the expected numbers of families for the Bayesian analysis, under the assumption that we learn about a randomly selected child from a randomly selected family. Thus, like Table 1, it assumes a mother is equally likely to mention a boy or girl if there is one of each. Similarly, the conditional probability that the mother mentions a girl named Florida, if she mentions a girl, equals the fraction of girls named Florida, α. Thus, for the 100 cases in which a BG family is selected, it is expected that a boy will be mentioned in 50 cases and a girl in 50 cases, with $\alpha 50$ of the mentioned girls named Florida and $(1 - \alpha)50$ not named Florida. For the GG column, no matter whether the mother mentions the older daughter or the younger daughter, the probability is α that the girl's name is Florida and $1 - \alpha$ that it is not Florida.

Table 4—Exact Expected Numbers of Families for Mlodinow's Question

	BG	GB	GG	Total
Families With a Girl Named Florida	$\alpha 100$	$\alpha 100$	$\alpha(2\text{-}\alpha)100$	$\alpha(4\text{-}\alpha)100$
Families With No Girl Named Florida	$(1\text{-}\alpha)100$	$(1\text{-}\alpha)100$	$(1\text{-}\alpha)^2 100$	$(3\text{-}\alpha)(1\text{-}\alpha)100$
Total	100	100	100	300

If the mother mentions a girl named Florida, F_m, the conditional probability that she comes from a two-girl family is one-half, regardless of the value of α:

$$P[\mathrm{GG} \mid \mathrm{F_m}] = \frac{\alpha 100}{\alpha 200} = \frac{1}{2}.$$

Similarly, if the mother mentions a girl not named Florida, N_m, the probability that she comes from a GG family is also one-half, regardless of the value of α:

$$P[\mathrm{GG} \mid \mathrm{N}_m] = \frac{(1-\alpha)100}{(1-\alpha)200} = \frac{1}{2}.$$

Thus, if the mother mentions a girl, the probability of two girls is one-half, no matter what her name is. The answer is the same as in our original Bayesian analysis: There is no paradox.

Mlodinow's Approach

Mlodinow does not answer the question he poses about the probability that a given family will have two girls, if we learn it has a girl named Florida. Instead, he gives an approximate answer to a different, hypothetical question: Among all BG, GB, and GG families that have a daughter named Florida, what proportion are GG? His approximation is one-half. This analysis parallels his take on the classic two-child problem.

Mlodinow purports to use a Bayesian approach, which he says "is to use new information to prune the sample space." However, a Bayesian analysis must account for the conditional probabilities that we would observe or learn about a girl, named Florida or otherwise, for different family types, and this Mlodinow does not do. Instead, he does a restricted sample space analysis much like the conventional analysis of the two-child problem. Despite these issues, his analysis is instructive.

Mlodinow assumes 100 million families, but we will assume 400 to be comparable with our earlier analyses. Let G_F indicate that a girl is named Florida and G_N that she is not. We will denote birth order as above.

With his assumption that α is tiny—one in a million—Mlodinow infers that the expected number of G_FG_F families (equal to $\alpha^2 100$) is, for practical purposes, zero. He also infers, since girls not named Florida are almost as numerous as boys, that the expected number of BG_F families (equal to $\alpha 100$) is, for practical purposes, equal to that of G_NG_F families (equal to $\alpha(1-\alpha)100$).

The commonality in these assumptions is that second-order terms in α^2 are so small they can be ignored. Since BG_F and G_FB families are expected in equal numbers, as are G_NG_F and G_FG_N families, Mlodinow then concludes that BG_F, G_FB, G_NG_F, and G_FG_N families are all equally likely. With G_FG_F families ruled out, the proportion of GG families in the BG, GB, and GG families with a girl named Florida is then equal to one-half.

To provide a further rationale for his analysis, Mlodinow offers a second set of calculations. Table 4 shows the relevant exact calculations of expected numbers of families in which there is at least one girl. We can rule out BB, and thus start with the 300 BG, GB, and GG families.

Of the 100 BG and 100 GB families, a fraction α has a girl named Florida. Of the 100 GG families, a fraction α of the older girls is named Florida and, of the $(1-\alpha)100$ families in which the older girl is not named Florida, a fraction α of the

younger girls is named Florida, giving a total of $\alpha 100 + (1 - \alpha)\, \alpha 100 = \alpha(2 - \alpha)100$ GG families with girls named Florida. The second row of the table can be most easily obtained by subtracting the entries in the first row from the totals for each column.

To translate his presentation to our framework, Mlodinow would observe that we expect 2 $\alpha 100$ families with a girl named Florida to be BG or GB, as shown in Table 4. He would then note that, of the 100 expected GG families, $\alpha 100$ will have an older child named Florida and $\alpha 100$ will have the younger child named Florida, yielding 2 $\alpha 100$ GG families with a girl named Florida.

The approximation $\alpha^2 = 0$ is important in this case as well, as is clear from the GG entry in the first row of Table 4, because otherwise there would be double counting of families in which both girls are named Florida. Mlodinow would then conclude that the fraction of GG families in BG, GB, and GG families with a girl named Florida is approximately equal to $2\,\alpha 100/(2\,\alpha 100 + 2\,\alpha 100) = 1/2$.

It turns out that this last set of calculations approximates an exact calculation in which girls are counted, rather than families. In BG and GB families combined, there are expected to be exactly 2 $\alpha 100$ girls named Florida, and the same number in GG families. There are half as many GG families as there are BG plus GB families combined, but twice as many girls per family and thus twice as many girls named Florida, under the independence assumption.

Thus, in the second variant of his analysis, Mlodinow is in effect answering yet another, but still hypothetical, question: What fraction of girls named Florida from two-child families is from two-girl families? However, the exact analysis of this case applies for a girl of any name, and indeed for girls in general. If a girl is randomly selected from all BG, GB, and GG families, the probability that she came from a GG family is exactly one-half. In particular, the fraction of girls named Florida from two-girl families is independent of the value of α: $2\,\alpha 100/(2\,\alpha 100 + 2\,\alpha 100) = 1/2$. Thus, it is irrelevant in this variant of the analysis that the girl has an unusual name.

Finally, we can generalize Mlodinow's analysis on its own terms, using the exact formulas in Table 4. For families with a girl named Florida, the fraction of BG, GB, and GG families that is GG equals $(2 - \alpha)/(4 - \alpha)$. Mlodinow, in effect, finds the limit of this formula as α approaches zero, obtaining his answer of one-half.

The other extreme is for α to approach one. All girls are named Florida, so that being a girl named Florida is equivalent to being a girl. In this case, the formula evaluates to one-third, which is the fraction of all BG, GB, and GG families that are GG.

In general, if we consider families with a girl named Florida, as the name becomes increasingly rare, the number of families that is GG goes down more slowly than the combined number of those that are BG, GB, or GG, and so the fraction of BG, GB, or GG families that are GG goes up.

Discussion

A general question is how best to accommodate new information into the evaluation of uncertain situations. Use of the restricted sample space approach for the two-child problem does not yield a proper conditional probability that a family has, say, two girls, given that one has learned that one of the children is a girl. All it offers, in this case, is a hypothetical calculation of the fraction of BG, GB, and GG families that are GG. In the classic two-child problem, it also offers an erroneous illusion of simplicity—that, in general, a two-child family is equally likely to be BG, GB, or GG if we learn one of the children is a girl.

In contrast, the Bayesian approach provides useful conditional probabilities that can be applied directly to a family at hand as we acquire new information about it. It also provides discipline in that it requires us to be clear about the full set of assumptions that enter into our probabilistic inferences.

In this analysis, we think we have made plausible assumptions about the probability that a mother would mention a girl versus a boy, depending on whether the family is BB, BG, GB, or GG, but the Bayesian approach offers the flexibility to accommodate alternative sets of assumptions in any case. C

Further Reading

Bar-Hillel, M., and R. Falk. 1982. Some teasers concerning conditional probabilities. *Cognition* 11:109–122.

Freund, J. E. 1965. Puzzle or paradox? *The American Statistician* 19:29–44.

Gardner, M. 1959. *The scientific American book of mathematical puzzles and diversions*. New York: Simon and Schuster.

Gardner, M. 1961. *The 2nd scientific American book of mathematical puzzles and diversions*. New York: Simon and Schuster.

Mlodinow, L. 2008. *The drunkard's walk: How randomness rules our lives*. New York: Pantheon Books.

Nickerson, R. S. 1996. Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin* 120:410–433.