Following is a letter to my son, Alexander

Alexander,

Recently, I told you about a waitress who asked me the probability of one of her sisters sharing their mother's birthday. I told her, and you, that, with four children, the probability is about .01, or about 1 in every 100 four-children families. You asked how I arrived at the answer, and I quickly went through the assumptions and method, but here is a more complete explanation.

First, to me, the interesting thing about probability problems is that, although you need some simple math to get a final numeric answer, their difficulty lies not in math but in the logic that defines the events being studied and, given the outcome in question, identifying which events of all those possible agree with a positive outcome (note that in probability, a positive outcome, or success, does not necessarily mean a preferred or pleasant outcome – when the question is the probability of a bee sting, a successful outcome is getting stung). Humans tend to use the least amount of information possible to reach confidence in an answer and we tend to revise our conclusions only after new data or circumstances are introduced. I think that most people would answer the following questions differently if asked in the order listed:

1. What is the probability of dying in a car accident?
2. What is the probability of dying in a car accident on Mt. Everest?

So the trick is to 1.) exhaustively identify all possible events that relate to the outcome, and 2.) identify those events that align or agree with our definition of success. Their ratio is the probability of success. The difficulty is in *exhaustive* (all events or possible conditions – if your question is the probability of a flat tire, do you separate trucks from cars?) and *relates* (some – most – events have no relationship to the outcome in question – Monday through Friday data should be excluded from the event population when asking the probability of weekend events). There's also the problem of extraneous information (if all the girls at a particular school have blonde hair and where red shoes then the probability of a randomly selected student being a girl is equal to the probability a randomly selected student being a blonde girl with red shoes) which generally complicates things for no benefit and should be excluded – but being certain that relevant event descriptors are not.

So, what is the probability of a mother and child sharing a birthday?

Assumptions:

1. The mother has four children, although the following method can be used for any number.
2. Neither the mother nor any of the children were born in a leap year. Accounting for February 29th complicates things since the probability of a person having this as his birthday depends on whether or not he was born in a leap year: 1/366 for a leap year, 0 otherwise. All other birthdays have probability of 1/366=.002733 in a leap year and 1/365=.002739 in other years – very little difference. So, our answer is exact for families with no leap year birthdays and very close for those with.
3. Birthdays have equal distribution across calendar days. That is, the probability of a randomly selected person having a particular birthday is the same as for any other day of the year. As previously stated, we will use 1/365. Although this assumption is known to be invalid (more people are born in North America during September and October than in other months), we adopt it for simplicity and assert no loss of generality.
4. None of the siblings are twins, triplets, or quadruplets since this would effectively remove one, two, or three children from the sample space and be identical to a three, two, or, one child problem. Note, however, the interesting related question of the probability of twins, or triplets and shared birthday. The only requirement in answering this is to redefine your events to align with the question: Probability of quads AND shared birthday in four children = probability of shared birthday with one child AND probability of quads (more on the product of probabilities later).
5. A person has one and only one birthday.

6. The resulting probability is identical for a son or daughter. The waitress asked about mother-daughter probabilities, but the question is identical to that of mother-son, or father-son, father-daughter probabilities. The first step in solving any problem is to establish the objective. The waitress asked in terms of mother-daughter birthdays, but (I have confidence) she really meant parent-child, being more general and what most people would be interested in.

A few standard probability rules and notation will be used:

- P(*event*) is the probability that *event* will occur.
- 1-P(*event*) is the probability that *event* will not occur.
- When two events are independent (the probability of one occurring is not dependent on the other occurring or not occurring), P($event_1$ and $event_2$) = P($event_1$) X P($event_2$). Consider rolling an unbiased die twice. P(rolling two 1s) = 1/6 X 1/6. In fact all possible pairs (1,1), (1,2) … (6,6) have equal probability of 1/6 X 1/6 = 1/36. This agrees with the fact that there are 36 possible pairs (6 for roll 1 X 6 for roll 2) with each pair as likely as another, giving P(a given pair) = 1/36.
- $q \in \{set\} \Longrightarrow q$ is a member of *set*. $q \notin \{set\} \Longrightarrow q$ is not a member of *set*.

Method:

1. Let M be the mother's birthday and B_1, B_2, B_3, and B_4 be the birthdays of the four children.
2. There are 364 equally probable days such that B_1<>M. So P(B_1<>M) = 364/365.
3. Similarly, there are 364 days such that B_2, B_3, and B_4 are different from M.
4. Given the assumption of independence (knowing the birthday of one child gives no knowledge of that of another, no twins, etc.), P(B_1<>M and B_2<>M and B_3<>M and B_4<>M) = 1/364*1/364*1/364*1/364 = $1/364^4$ = .9891.
5. We just calculated P(all four children birthdays different than that of their mother) so P(at least one child shares his mother's birthday) = 1 – that value = 1 - .9891 = .0109 ≈ .01.

What if she had asked the probability of at least two family members sharing their birthday? (By the way, no one ever asks, "What is the probability of a certain event?" but "What are the odds or chances or how rare is such and such event?" We are comfortable with the notion of chance but, often, struggle to develop a good understanding of the events and relationships that lead to a reliable, likely, statement of outcome.) Let's include dad here, so we have five birthdays. Let them be B_1 through B_5. Then:

1. P(B_1<>B_2) = 364/365.
2. P($B_3 \in \{B_1, B_2\}$) = 363/365, since there are 363 days unequal to the first two birthdays.
3. Continuing, P($B_4 \in \{B_1, B_2, B_3\}$) = 362/365.
4. And, finally, P($B_5 \in \{B_1, B_2, B_3, B_3\}$) = 361/365.

Therefore, P(all 5 birthdays being different) = the product of 1 though 4, above, which is .97, making P(at least two members sharing a birthday) = .03, or 3%

Following is an interesting variation of this problem.

How many people must be present at a gathering to have confidence that at least two share the same birthday?

Assumptions:

1. No leap year birthdays, as before.
2. Birthdays have equal distribution across calendar days, as before.
3. With respect to birthday, the gathering consists of a random collection of individuals. No Twin, New Year's Baby, or St. Valentine's Day Birthday club meetings, etc.
4. A person has one and only one birthday.

Method 1:

Given a meeting of N people,

1. Randomly select two people, A_1 and A_2, with birthdays B_1 and B_2. $P(B_2<>B_1) = 364/365 = .9973$, since there are 364 possible days B_2 that are unequal to B_1 (note how this agrees with $1-P(B_2=B_1) = 1-1/365 = 364/365$).
2. Randomly select a third person, A_3, different than A_1 and A_2, with birthday B_3. Then $P(B_2<>B_1$ and $B_3 \notin \{ B_1, B_2\}) = P(B_2<>B_1) X P(B_3 \notin \{ B_1, B_2\}) = 364/365 X 363/365$, since there are 363 days unequal to B_1 and B_2. The multiplication is due to independent selection of people. So, after randomly selecting three people, P(all three having distinct birthdays) = $364*363/365^2 = .9918$.
3. Continuing this way, P(*n* people having distinct birthdays) = $364*363*…*(365-n+1)/365^{n-1}$.
4. But the question is P(at least one shared birthday among *n* people, $ALOSB_n$), which is 1-P(no shared birthdays, NSB_n) = $1-364*363*…*(365-n+1)/365^{n-1}$. So, for instance, $P(ALOSB_{10}) = 1-364*363*362*361*360*359*358*357*356/365^9 = 1-.8831 = .1169$.
5. Find *n* such that $P(ALOSB_n)$ = your desired level of confidence. Say you want to be 50% = .5 confident that at least two of *n* people will have a common birthday. Then $P(ALOSB_n) = 1 – P(NSB_n) = .5$, which however easy to write is a bit complicated to calculate. In fact, I have tried to derive a closed form equation (a one line function of *n* and .5 that can be entered into a calculator) and even did a bit of research, but haven't yet found anything. The problem is with the product of successive integers in the numerator of $P(NSB_n)$, that is $364*363*…*(365-n+1)$. There are interesting formulas for sums of consecutive integers (I have two of my own if you are interested, also for sums of squared and cubed integers), but not for products. Maybe you can find one and get (y)our name attached to a famous theorem. We do have $364*363*…*(365-n+1) = 364!/(365-n)!$ (remember, $k! = k*(k-1)*(k-2)*…*1$), but this still involves the product of successive integers. So much for closed form. But we can write an iterative algorithm to calculate $P(NSB_n)$ for *n* beginning at 1 and continuing with 2, 3, and so on until $1 - P(NSB_n) = .5$, something like:

```
Let n=1
Let p=(365-n)/365
While p>.5
  Let n=n+1
  Let p=p*(365-n)/365
End
Print n+1
```

The *n*+1 is due to starting the comparison with person two - we must add one for the first person.

Following are some sample results:

Desired Confidence	People Required
25%	15
50%	23
75%	32
90%	41
99%	57

Another way to interpret P(ALOSB$_n$) is, “If *n* people are at a gathering, what is the probability that at least two have a common birthday?” This is the gambler’s, or casino’s, approach – make a wager only if the probability of winning is in your favor, .5 and up for a winning record, .6 and up to win 60% of the time, etc. If you could repeatedly wager, at gatherings of 32 people, that at least two people share a birthday, you would win 75% of the time (if you require a 75% winning record then do not wager with fewer than 32 people present). Similarly, at gatherings of 15 people, a wager that no two people have a common birthday would win 75% of the time (or lose and cost you 25% of the time). If you took the casino approach and made one thousand, five dollar bets for ALOSB$_{25}$, you would expect to win 55%, or 550 of the wagers, while losing 45%, or 450, for earnings of 550*5 – 450*5 = 100*5 = $500, before expenses.

Method 2 (pairs of birthdays):

This method compares the birthdays of all possible pairs of people present at a gathering. Letting the number of people present be *n*, we have:

1. The total number of unique pairs of people is *n**(*n*-1)/2 (imagine having two columns, you can place n people in column 1, leaving *n*-1 for column 2, hence *n**(*n*-1), but, for instance, person 5 in column 1 and 8 in column 2 is the same pair as person 8 in column 1 and 5 in column 2, so there are two ways to represent each unique pair, hence the division by 2).
2. P(two people, one pair, having different birthdays) = 364/365.
3. P(all pairs of people having different birthdays) = 364/365 times itself once for each pair = (364/365)$^{\text{number\_of\_pairs}}$ = (364/365)$^{n*(n-1)/2}$ (due to the probability of a sequence of independent events being the product of the probabilities of the individual events, as before).
4. P(at least one pair with same birthday) = 1- P(all pairs different) = 1-(364/365)$^{n*(n-1)/2}$.

The last expression [P(at least one pair with same birthday) = 1-(364/365)$^{n*(n-1)/2}$] is the sought after, single line, closed form expression in terms of *n* (remember you want to wager knowing probabilities given *n*). So, given a desired probability of 75% that at least one pair, two people, share birthdays, this equation tells us

$$.75 = 1 - \left(\frac{364}{365}\right)^{n*(n-1)/2} \quad \Longrightarrow \quad \left(\frac{364}{365}\right)^{n*(n-1)/2} = 1 - .75 = .25 \quad \Longrightarrow$$

$$\frac{n^2 - n}{2} \log\left(\frac{364}{365}\right) = \log(.25) \quad \Longrightarrow \quad n^2 - n = 2\frac{\log(.25)}{\log\left(\frac{364}{365}\right)} \quad \Longrightarrow$$

$$n^2 - n - 2\frac{\log(.25)}{\log\left(\frac{364}{365}\right)} = 0.$$

The final equation is in standard quadratic form, making it a candidate for the quadratic formula (remember that from Algebra class), giving

$$n = \frac{1 + \sqrt{1 + 8\frac{\log(.25)}{\log\left(\frac{364}{365}\right)}}}{2} \quad = \quad 32.3$$

or 32 people, which agrees with our value of *n* from method 1 for confidence=.75.

Imagine counting party attendees and quickly entering this formula into your i-phone TI-89 app (oops, I don’t think it’s available, you’ll have to borrow someone’s android) then raking in the bucks.

Method 3 (pairs, different algebra, and a loop):

Again we use pairs of people but, this time, instead of asking the probability of all pairs having different birthdays (then subtracting from 1), we ask the probability of pair 1 OR pair 2 OR … OR pair *n* having the same birthday. For mutually exclusive outcomes, such as rolling a 1 or a 6 in a single roll of a fair die, the probability of outcome A OR outcome B is P(A) + P(B). In the case of the die roll, P(1 or 6) = P(1) + P(6) = 1/6 + 1/6 = 1/3, which makes sense since 1 and 6 compose 1/3 of the possible outcomes. But, it gets interesting with multiple trials, such as the probability of rolling a 1 or 6 in two attempts. Here, of the 36 possible outcomes (1 through 6 for roll 1 X 1 through 6 for roll 2), there are 20 pairs (11, 12, 13, 14, 15, 16, 61, 62, 63, 64, 65, 66, 21, 31, 41, 51, 26, 36, 46, and 56) with a 1 or 6 in one of the rolls, making P(1 or 6) = 20/36 = 5/9. Because the outcome of roll 1 neither excludes nor predicts the outcome of roll 2, we can't simply add probabilities [P(1 or 6 in roll 1)=1/3 and P(1 or 6 in roll 2)=1/3, but P(1 or 6 in roll 1 OR 1 or 6 in roll 2) is not 1/3+1/3=2/3, as we have already demonstrated]. The general rule for P($event_1$ or $event_2$), is developed as follows:

> Let there be a or trial with possible outcomes A and A_N, where A_N indicates NOT A (an outcome of A_N means that A was not the outcome). Then, in a single trial, P(A OR A_N) = 1, since one or the other must result. Now in two trials, the possible paired results are AA, AA_N, A_NA, and A_NA_N. These are mutually exclusive (one and only one pair must result) and three have A as an outcome (AA, AA_N, and A_NA). Therefore,
> P(A in trial one or A in trial 2) = P(AA) + P(AA_N) + P(A_NA). But, since the trials are independent,
>
> P(AA) = P(A)*P(A), P(AA_N) = P(A)*P(A_N), and P(A_NA) = P(A_N)*P(A), giving
>
> P(A in trial 1 or 2) = P(A)*P(A) + P(A)*P(A_N) + P(A_N)*P(A) = P(A)*[P(A)+P(A_N)] + P(A_N)*P(A) =
>
> P(A)*(1) + P(A_N)*P(A) = P(A) + P(A_N)*P(A).
>
> But, P(A_N)*P(A) = P(A)*P(A_N) = P(A)*[P(A) + P(A_N)] - P(A)*P(A) = P(A) - P(A)*P(A), so
>
> P(A in trial 1 or 2) = P(A) + P(A) - P(A)*P(A) = 2P(A) – P(A)^2.

This generalizes to P($event_1$ or $event_2$) = P($event_1$) + P($event_2$) - P($event_1$)*P($event_2$), using the same algebra as above.

Simple, huh? So, getting back to the probability of pairs of people having a common birthday, suppose there are *n* people present, giving *n**(*n*-1)/2 unique pairs of people, as before. Then
P(pair *i* has common birthday, PCB_i) = 1/365, giving

> P(PCB_1 or PCB_2) = P(PCB_1) + P(PCB_2) - P(PCB_1)*P(PCB_2) = 1/365 + 1/365 - (1/365)^2 = .005472.

Now, think of PCB_1 or PCB_2 as $event_1$. Then

> P(PCB_1 or PCB_2 or PCB_3) = P($event_1$ or PCB_3) = P($event_1$) + P(PCB_3) - P($event_1$)*P(PCB_3) =
>
> .005472 + 1/365 - .005472*(1/365) = .008197.

Notice how the probability of a shared birthday increases as the number of pairs increases, from 1/365 = .002740 for one pair, to .005472 for two pairs, to .008197 for three pairs. This is expected since the more pairs you observe, the more you would expect to find a certain trait. Continuing,

> P(PCB_1 or PCB_2 or PCB_3 or PCB_4) = .008197 + 1/365 - .008197*(1/365) = .010914, and

P(PCB$_1$ or PCB$_2$ or PCB$_3$ or PCB$_4$ or PCB$_5$) = . 010914 + 1/365 - . 010914*(1/365) = .013624.

So, we simply need to continue adding pairs until our desired level of confidence is met. Here is a simple process for accomplishing this (using .75 confidence):

```
Let pairs=1
Let p=1/365
While p<.75
  Let pairs=pairs+1
  Let p=p+1/365-p/365
End
Print pairs
```

I programmed this on a real TI-84 (that I searched for and found in some of your belongings) and got the answer `pairs` = 506 (after about 10 seconds of run time).

To get the number of people required, *n*, recall that pairs = *n**(*n*-1)/2, giving

$$\frac{n^2-n}{2} = 506 \implies n^2 - n - 1012 = 0 \implies \frac{1+\sqrt{1+4048}}{2} = 32.3,$$

(the quadratic formula again), which is the same answer as in methods 1 and 2.

But wait, there's more!

We saw that P(*event*$_1$ or *event*$_2$) = P(*event*$_1$) + P(*event*$_2$) - P(*event*$_1$)*P(*event*$_2$), so letting
p = P(pair i has common birthday, PCB$_i$) = 1/365 for every pair, and $q = 1 – p$, we have

P(PCB$_1$ or PCB$_2$) = $p + p - p*p = p*(p\text{-}1) + p = p*q + p \Longrightarrow$

P(PCB$_1$ or PCB$_2$ or PCB$_3$) = $(p*q + p) * (p\text{-}1) + p = (p*q + p)*q + p = p*q + p*q^2 + p \Longrightarrow$

P(PCB$_1$ or PCB$_2$ or PCB$_3$ or PCB$_4$) = $(p*q + p*q^2 + p)*q + p = p*q + p*q^2 + p*q^3 + p$,

and the pattern for k pairs is

P(PCB$_1$ or PCB$_2$ or PCB$_3$ or PCB$_4$ or … or PCB$_k$) = $p + p*q + p*q^2 + p*q^3 + \ldots + p*q^{k\text{-}1}$.

Noting that $p = p*q^0$, we have

P(PCB$_1$ or PCB$_2$ or PCB$_3$ or PCB$_4$ or … or PCB$_k$) = $p\sum_{i=0}^{k-1} q^i$,

which is *p* times a truncated geometric series (do you recall this famous Taylor series from calculus?). Noting that, for 0<*q*<1, the Taylor series gives

$$\frac{1}{1-q} = \sum_{i=0}^{\infty} q^i \implies \sum_{i=k}^{\infty} q^i = \frac{q^k}{1-q} \implies \frac{1}{1-q} = \sum_{i=0}^{k-1} q^i + \sum_{i=k}^{\infty} q^i \implies$$

$$\sum_{i=0}^{k-1} q^i = \frac{1}{1-q} - \sum_{i=k}^{\infty} q^i = \frac{1}{1-q} - \frac{q^k}{1-q} = \frac{1-q^k}{1-q}$$

Now, multiplying this by *p* (which is 1-*q*), we have

P(PCB_1 or PCB_2 or PCB_3 or PCB_4 or … or PCB_k) = $p\frac{1-q^k}{1-q} = (1-q)\frac{1-q^k}{1-q} = 1 - q^k$.
And, recalling that *p* = 1/365, this becomes 1 – $(364/365)^k$.

Amazing! Another closed for solution. Well, actually, it's the same one we arrived at in method 2, but from a completely different and independent approach. Don't forget that *k*, here, is the number of pairs. If our desired success confidence level is .75, then the number of pairs required is found by solving

$$.75 = 1 - \left(\frac{364}{365}\right)^k \;\;\Rightarrow\;\; k = \frac{log\,(.25)}{log\left(\frac{364}{365}\right)} = 505.3$$

Look familiar? To get the number of people, *n*, recall that *k* = *n**(*n*-1)/2 ⟹ $n^2 – n$ - 1010.6 = 0, or (from the quadratic formula) *n* = (1+√(1+4042.4))/2 = 32.3, the same answer as in all of the previous methods.

Please forgive me, but I thought of yet another approach to a solution. And so,

Method 4 (the geometric probability approach):

The geometric is one of the most intuitive of the common probability distributions. It measures the probability of a successful event after *k* non-successes (the events, or trials are assumed to be independent). Given the probability, *p*, of success in a single trial, the probability of non-success is 1-*p* and, due to independence of trials, P(*k* non-successes followed by a single success) = $p(1\text{-}p)^k$. It so happens that the expected value of *k* (the average number of trials you would expect before observing a successful one) is 1/*p* (there is some real interesting math that establishes this, moment generating functions and integration by parts, that I will omit here so that I can finish this letter). Letting a trial be the sampling of a pair of people and testing for common birthday, with probability of success *p* = 1/365 in any one trial, the expected number of trials before success is 1/*p* = 1/(1/365) = 365. So, in repeated gatherings of people, adding one pair at a time until observing a pair with common birthday, the average number of pairs would be 365, giving an average party size *n*, from the quadratic formula, of

$$\frac{n^2 - n}{2} = 365 \;\;\Rightarrow\;\; n^2 - n - 730 = 0 \;\;\Rightarrow\;\; n = \frac{1 + \sqrt{1 + 2920}}{2} = 27.5$$

And so, there you have it … or them. Four answers, plus a variation (in method 3), to your question. I hope you enjoy them.

Dad