TURNING TECHNOLOGY LOOSE ON VARIATIONS OF THE BIRTHDAY PROBLEM

Dr. Joel C. Fowler
Mathematics Department, Southern Polytechnic State University
1100 South Marietta Parkway
Marietta, GA 30060-2896
jfowler@spsu.edu

Abstract: The classic Birthday Problem finds the probability that at least two people in a group of k share the same birthday. The solution involves nothing more than basic counting and probability. Here we look at extensions of this question with solutions that are not as elementary, such as finding the probability that at least m people in a group of k share a common birthday, for values of m larger than 2. We also examine the average number of days on which there are at least m common birthdays, and the likelihood of more than a single birthday being shared by a group of people. We compare the exact values found for some of these problems with results from simpler approximation methods, such as the Poisson distribution. These generalizations require substantially more theory and computational power than the traditional birthday problem. The solutions involve recurrence relations based on discrete probability computations and finding means using indicator random variables. Their complexity makes technology essential in implementing the necessary computations. The solutions are at the level of a senior undergraduate mathematics major special project. The technology used is Maple.

The classic Birthday Problem from elementary probability is to find the likelihood that at least two people in a group of k share a common birthday, and determine the smallest value of k for which that probability exceeds .5 . Its solution is a nice application of complements and elementary counting techniques. The probability is given by:

$$1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - k + 1)}{365^k} ,$$

since this is simply the complement of the event that all k birthdays fall on different days. And it is well known that k = 23 is the turning point at which the probability exceeds .5 . Many people are surprised at how low that threshold is.

Here we are concerned with finding answers to more complicated related questions with less elementary solutions. For example, for a group of k people, one might ask about the probability of at least one day with m common birthdays, for m > 2, or the likelihood of d days with a common birthday for d > 1, or the expected number of days with more than one birthday. The solutions to these questions are more involved mathematically and computationally.

In order to approach these questions, we define:

Prob(k, n, m, d) = the probability of at least d days, each with at least m birthdays,
in an n day year, for a group of k people.

We first find a computationally efficient recursion for these probabilities for the case when d = 1. That is, we will be working with the probability that at least d=1 day in an n day year has at least m birthdays, for a group of k people. The recursion will induct on the number of days in the year, n.

We break our computation of Prob(k, n, m, 1) into cases based on the number of birthdays that occur on the nth day of the year. If we let i be that number, then we have for Prob(k, n, m, 1):

$$\sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}\,\mathrm{Prob}(k-i, n-1, m, 1) + \left(1 - \sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}\right),$$

with the last term being the probability that at least m birthdays occur on the nth day. With some rearrangement we then have the working formula:

$$\mathrm{Prob}(k, n, m, 1) = 1 - \sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k}\,(1 - \mathrm{Prob}(k-i, n-1, m, 1)),$$

with the initial conditions that Prob(k, n, m, 1) = 0 if k < m or n = 0.

The following Maple code evaluates these probabilities for a range of values and stores them. Note that it takes advantage of simpler direct computations, from the traditional birthday problem, for m=2.

```
> Maxn := 365 ; Maxk := 400 ; Maxm := 5:
> st := time():
     for n from 0 to Maxn do  for k from 0 to Maxk do if n=0 then Prob(k,n,2,1):=0 else
    Prob(k,n,2,1) := evalf( 1 - binomial(n,k)*k!/(n^k) ) end if; end do; end do;
      for m from 3 to Maxm do
       for n from 1 to Maxn do
        for k from 1 to Maxk do
       if k>(m-1)*n then Prob(k,n,m,1) := 1 else if k<m then Prob(k,n,m,1) := 0 else
       Prob(k,n,m,1):= evalf( 1 - add( ( binomial(k,i)*(n-1)^(k-i)/n^k
       )*(1-Prob(k-i,n-1,m,1)) ,   i=0..(m-1) ) ) end if end if;
    end do; end do; end do;
time() - st;
```

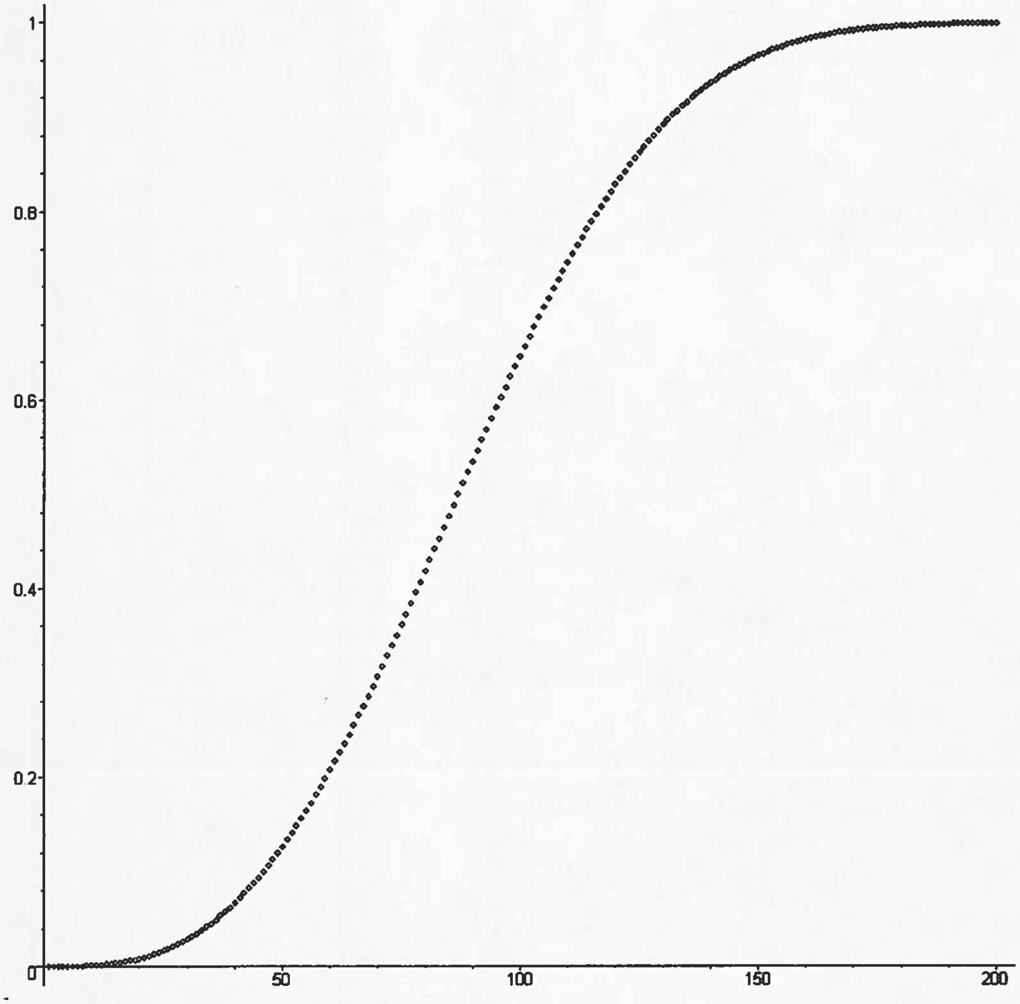
This routine took less than 2 minutes to run. With the probabilities stored we can examine how the probability changes for n=365 and various values of m and k.

For example, if one is concerned with having at least one triple birthday, we evaluate:

```
> seq([10*k,Prob(10*k,365,3,1)],k=1..15);
>   [10, 0.0008877457], [20, 0.0082426865], [30, 0.0285305042],
    [40, 0.0668894861], [50, 0.1263751846], [60, 0.2072303207],
    [70, 0.3064873273], [80, 0.4181689338], [90, 0.5341955714],
    [100, 0.6458645064], [110, 0.7455259085], [120, 0.8279641283],
    [130, 0.8910760072], [140, 0.9356999330], [150, 0.9647665029]
```

We can see the trend using pointplot:

```
> with(plots):  pointplot({seq([k,Prob(k,365,3,1)],k=1..200)});
```



The turning point for a better than even chance of a triple birthday is between k = 80 and k = 90 people. With a little trial and error we find it exactly as k = 88.

A similar trial and error process reveals the turning points for m = 3, 4, and 5. Thus we have the following for m from 2 through 5:

```
>   [23,2,Prob(23,365,2,1)];[88,3,Prob(88,365,3,1)];[187,4,Prob(187,365,4,1)];
[313,5,Prob(313,365,5,1)];
>                 [23, 2, 0.5072972343]
                  [88, 3, 0.5110651107]
                  [187, 4, 0.5026853730]
                  [313, 5, 0.5010704762]
```

We now turn our attention to the likelihood of more than one day with at least m birthdays. That is, d > 1. We continue to use the approach we've followed by inducting on the number of days in the year. Unfortunately the recurrence is much deeper because we must involve the number of people having a birthday on the nth day of the year, no matter how large that is. Previously, because we were only interested in at least one day with m, we could simply stop once that number was at least m and combine those cases into a single term. Now, since we wish more than a single day with at least m, we must retain the knowledge of exactly how many people are remaining to distribute onto the other days. That is, we were previously able to employ an order m recurrence, regardless of k. Now it must be order k.

Reasoning as before, where i = 1, 2, . . . , k is the number of people with a birthday on the nth day, we have:

$$\begin{aligned}\mathrm{Prob}(k,\ n,\ m,\ d)\ &=\ \sum_{i=0}^{m-1}\frac{\binom{k}{i}\,(n-1)^{k-i}}{n^{k}}\,\mathrm{Prob}(k-i, n-1, m, d)\\ &+\sum_{i=m}^{k}\frac{\binom{k}{i}\,(n-1)^{k-i}}{n^{k}}\,\mathrm{Prob}(k-i, n-1, m, d-1)\ ,\end{aligned}$$

with the initial conditions that Prob(k, n, m, d) = 0 whenever k < md or d > n.

To avoid long runtimes, we implement this recurrence in Maple for m = 2 only. That is, we will be finding the probability of at least d days with more than one birthday. Although this is limiting, it is the most natural next birthday question after finding probabilities for at least one day with multiple common birthdays. The following routine, which assumes the previously computed values for m = 1 and took approximately 50 minutes to run, yields probabilities for a range of values with m = 2.

```
> Maxd := 15: Maxn := 365: Maxk := 200:
> st := time() :
      for d from 6 to Maxd do
      for n from 0 to Maxn do
      for k from 0 to Maxk do
          if k<2*d then Prob(k,n,2,d) := 0 else if n=0 then Prob(k,n,2,d) := 0 else
          Prob(k,n,2,d):= evalf( (((n-1)/n)^k)*Prob(k,n-1,2,d)+
           ((k*(n-1)^(k-1))/(n^k))*Prob(k-1,n-1,2,d) +
          add( ( binomial(k,i)*(n-1)^(k-i)/n^k    )*Prob(k-i,n-1,2,d-1), i=2..k ) )
        end if;  end if;
      end do; end do; end do;
time() - st;
```

From this computation we find that the chances of finding at least two days with at least two birthdays in a group of 23 is:

```
> Prob(23, 365, 2, 2);
                              0.1363714950
```

In a group of 50 people this same event has probability:

```
> Prob(50, 365, 2, 2);
                              0.8495070001
```

This is quite likely. In fact, going further with k = 50 and larger numbers of days with at least 2 birthdays, we find:

```
> Prob(50, 365, 2, 3);
                              0.6242529504
> Prob(50, 365, 2, 4);
                              0.3694489176
```

These are surprisingly high probabilities. In recalling the original birthday problem one is motivated to ask what the 50% turning point is for d days with at least two birthdays, for d = 1, 2, . . . 10. Some graphing, together with trial and error, yields the following thresholds:

```
> [1,23,Prob(23,365,2,1)]; [2,36,Prob(36,365,2,2)]; [3,46,(46,365,2,3)];
[4,55,Prob(55,365,2,4)]; [5,62,Prob(62,365,2,5)]; [6,69,Prob(69,365,2,6)];
[7,75,Prob(75,365,2,7)]; [8,81,Prob(81,365,2,8)]; [9,86,Prob(86,365,2,9)];
[10,92,Prob(92,365,2,10)];

                         [1, 23, 0.5072972343]
                         [2, 36, 0.5005548853]
                         [3, 46, 0.5037252118]
                         [4, 55, 0.5247976698]
```

```
[5, 62, 0.5136663761]
[6, 69, 0.5232846131]
[7, 75, 0.5170479762]
[8, 81, 0.5229263173]
[9, 86, 0.5069590963]
[10, 92, 0.5309244813]
```

So , for example, if one would like a better than even chance of at least 5 different days with more than one birthday, at least 62 people are required.

These surprisingly low values of k are supported by examining the expected number of days in an n day year that have at least m birthdays, for a group of k people. To attack this problem we define:

E(k, n, m) = the expected number of days with at least m birthdays in an n day year for a group of k people.

A relatively simple use of indicator random variables provides a direct computation of E(k, n, m). For i = 1, 2, 3, . . ., n we let

$$X_i = \begin{cases} 0, & \text{if there are fewer than m birthdays on day i} \\ 1, & \text{if there are at least m birthdays on day i} \end{cases} .$$

Then clearly the number of days with at least m birthdays is the sum of these random variables. Counting the number of ways to have exactly i birthdays on a given day is straightforward. Hence we may easily compute the expected value of each of these random variables as:

$$E(X_i) = 0 \cdot \sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} + 1 \cdot \left(1 - \sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} \right) .$$

Since the expected value of a sum is the sum of the expected values, and there are n of these random variables, we have:

$$E(k, n, m) = n \left(1 - \sum_{i=0}^{m-1} \frac{\binom{k}{i}(n-1)^{k-i}}{n^k} \right)$$

Since this expression demands very little computing power, it provides a nice shortcut to getting a sense of how probabilities rise with k. And evaluating this for various values of k and m with n = 365 reveals that the expected number of days grows rather quickly as k increases. For example, we have for the expected number of days with at least 2 birthdays:

E(40, 365, 2) = 1.994
E(60, 365, 2) = 4.364
E(80, 365, 2) = 7.516
E(100, 365, 2) = 11.360
E(120, 365, 2) = 15.812
E(140, 365, 2) = 20.797

This is certainly in line with our earlier discovery that, for a better than even chance of 10 days with at least 2 birthdays, only 92 people are needed.

We close with the looser question as to whether there is "magic number" for extensions of the Birthday Problem that parallels the k = 23 answer for the traditional problem. Certainly, given the variety of questions that could be asked concerning the number of days and/or number of birthdays on a day, there is no one complete answer. The work we have done, however, does point to one simple threshold for the sorts of extensions that most naturally come to mind.

Recall that k = 88 is the turning point for a better than even chance of having at least one day with at least three birthdays. Noting how the likelihood of multiple days with at least two birthdays grows, we see that, for k = 88:

```
> seq([i,Prob(88, 365, 2, i) ] , i=1..10 );
[1, 0.9999892802], [2, 0.9998252804], [3, 0.9986388419],[4, 0.9932325040],
[5, 0.9757937035], [6, 0.9333674275],[7, 0.8523683644], [8, 0.7277018879],
[9, 0.5700024362], [10, 0.4036773712]
```

Thus we find that, with 88 people you . . . .
. . . . have a better than even chance of a triple (or more) birthday,
. . . . have a better than even chance of at least 9 days with more than one birthday,
. . . . are extremely likely to have at least 5 days with more than one birthday.

Put informally, 88 people might be a reasonable "magic number" for commonly thought of birthday coincidences that go beyond the standard Birthday Problem. That is, a room of 88 people is more likely than not to be teaming with all sorts of birthday coincidences. Finding shared birthdays is sometimes used as an icebreaker for larger groups of people. This result could be interpreted to indicate that such an icebreaker will be most successful, in terms of generating interesting coincidences, for groups of 88 or more.

As a check, we can easily simulate birthdays in Maple and check the distribution of coincidences. The following procedure generates a random collection of k birthdays, tallies how many days have exactly 1 birthday, 2 birthdays, 3 birthdays, etc. and reports the results in a vector form.

```
> with(RandomTools): with(Statistics):
> SimulationDistribution := proc(k::integer)
     Test := Vector(k, Generate(integer(range=1..365), makeproc=true)):
     Tempo := Tally(Test): Length:= Count(Tempo):
     Freqs := sort( [ seq(op(2,Tempo[i]),i=1..Length) ]  );
  Tally(Freqs);
end proc;
```

For example,

```
> SimulationDistribution(100);
                    [1 = 77, 2 = 10, 3 = 1]
```

indicates that for this sample of 100 birthdays, 77 days had a single birthday, 10 days had two birthdays, and 1 day had three. To examine k = 88, we create 20 separate samples.

```
> for i from 1 to 20 do SimulationDistribution(88); end do;
                    [1 = 71, 2 = 7, 3 = 1]
                    [1 = 71, 2 = 7, 3 = 1]
                    [1 = 70, 2 = 9]
                    [1 = 82, 2 = 3]
                    [1 = 72, 2 = 8]
                    [1 = 70, 2 = 9]
                    [1 = 66, 2 = 8, 3 = 2]
                    [1 = 62, 2 = 10, 3 = 2]
                    [1 = 64, 2 = 9, 3 = 2]
                    [1 = 72, 2 = 8]
                    [1 = 69, 2 = 8, 3 = 1]
                    [1 = 67, 2 = 9, 3 = 1]
                    [1 = 64, 2 = 12]
                    [1 = 62, 2 = 13]
                    [1 = 74, 2 = 7]
                    [1 = 61, 2 = 12, 3 = 1]
                    [1 = 60, 2 = 14]
                    [1 = 68, 2 = 10]
                    [1 = 70, 2 = 9]
                    [1 = 74, 2 = 7]
```

Thus we find that 8/20 of the samples had at least one triple birthday. While not 50%, this is within a reasonable margin of error. And we have 12/20 with at least nine days having more than one birthday. This, too, is in line with our expectation, as is the finding that all but one of the samples have at least 5 days with multiple birthdays.