THE BIRTHDAY PROBLEM AND GENERALIZATIONS

TREVOR FISHER, DEREK FUNK AND RACHEL SAMS

1. Introduction

The question that we began our comps process with, the Birthday Problem, is a relatively basic problem explored in elementary probability courses. To solve it, we find the probability that in a group of n people, two of them share the same birthday. The reason this problem is intriguing is that the probability values that we get as a result of our solution are much different that what one may expect. For example, in a room of size 23, the probability is already $\frac{1}{2}$ that two people share a birthday.

2. Basic Birthday Problem

The basic birthday problem is as follows:

> What is the probability that, in a group of n people, two of them share the same birthday?

If A is the event that two people share the same birthday, $\overline{A}$ is the event that no two people share the same birthday. Since $P(\overline{A})$ is easier to solve for than $P(A)$, we solve for $P(A) = 1 - P(\overline{A})$.

We can write $P(\overline{A})$ as the product of conditional events, where each event is a person having their birthday on a day not "occupied" by any previous person's birthday. Inherent in this line of thought is the assumption that all birthdays are equally likely, with the probability of each birthday being $\frac{1}{365}$. In other words,

$$\begin{aligned} P(A) &= 1 - P(\overline{A}) \\ &= 1 - \left(\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365-(n-1)}{365}\right) \\ &= 1 - \left(\frac{365 \cdot 364 \cdot \ldots \cdot (365-n+1)}{365^n}\right) \end{aligned}$$

$$= 1 - \frac{365!}{(365-n)!365^n}$$

However, the classical statement of the birthday problem is as follows:

> What is the smallest number of people required so that there is probability $\frac{1}{2}$ that two people share a birthday?

To answer this question, we look at values given by the above function for $P(A)$:

n	p(n)
10	.117
20	.411
22	.476
23	.507
30	.706
57	.99

We see that probability $\frac{1}{2}$ is bounded most closely above by $P(n = 23) = .507$, so 23 is the size group we need to have probability $\frac{1}{2}$ of having two people share the same birthday.

This problem, though seemingly simple, leads to several much more complex generalizations; the first of which is called the "Almost Birthday Problem".

3. Almost Birthday Problem

The Almost Birthday Problem is the simplest generalization of the Basic Birthday Problem, and is as follows:

> What is the probability that, in a group of n people, two of them have birthdays in the same interval of k days?

We approach this problem with the same method as the Basic Birthday Problem, by dealing with the complement of the event that we wish to examine.

First, we define A_k as the event that two people's birthdays are within k days of each other. Thus, the complement of A_k, which we write as $\overline{A_k}$, is the event that no two birthdays lie within the same interval of k days. As before, by the definition of conditional probability we have:

$$P(A_k) = 1 - P(\overline{A_k}) = 1 - P(\overline{A_k} \mid \overline{A_1})P(\overline{A_1})$$

In this equation, the event $\overline{A_1}$ is the event that no two people's birthdays are within the same interval of 1 day, or put more simply that no two people's birthdays coincide. By conditioning $\overline{A_k}$ on $\overline{A_1}$ and preventing coincidental birthdays, we simplify the system of ordering later in the solution. Thus, by the conclusion of Section 1,

$$P(\overline{A_1}) = \frac{365!}{(365-n)!365^n}$$

In order to approach the probability of two birthdays being in the same interval of k days, we must first find the number of possible orderings of birthdays that satisfy the conditions of our complement: that no two birthdays lie in the same interval of k days.

To create the orderings, we will define an indicator variable, I_j, with j indexing the days of the year (with January 1 represented as $j = 1$ and December 31 represented as $j = 365$). For each day of the year, I_j will give us a value of 1 if day j is someone's birthday, or a value or 0 if day j is not someone's birthday. Recall that no day can contain two birthdays, due to our conditional probability. We write I_j in the following way:

$$I_j = \begin{cases} 1, & \text{if there is a birthday on day} j \\ 0, & \text{else} \end{cases}$$

To visualize a potential ordering of birthdays that satisfies our conditions for $\overline{A_k}$, we imagine all of the values for I_j written out in a line. It would appear as follows:

$$1, \underbrace{0, 0, \ldots, 0}_{k-1}, 1, \underbrace{0, 0, \ldots, 0}_{k-1}, 1, \underbrace{0, 0, \ldots, 0}_{k-1}, 0*, 0*, \ldots$$

Note that 1's represent birthdays, un-starred 0's represent the first $k-1$ non-birthday days after a birthday, and starred 0's represent extra non-birthday days after the first $k-1$.

Now, imagine that we pull this line of 1's, 0's and 0∗'s into a circle and fix the first birthday that happens in the year. We do this so that intervals that contain both of the tail ends of the year don't get cut off on either end. For example, if our interval $k = 5$, and it were centered around January 1st, we would want to catch birthdays from December 30th through January 3rd.

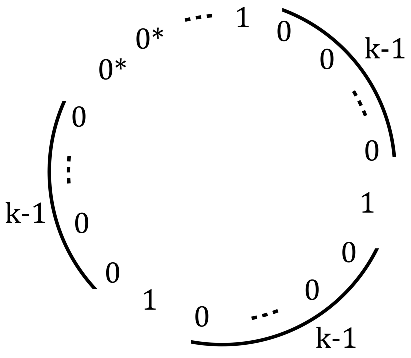

Figure 1. A possible circular ordering of birthdays

In order to find the probability of not having any birthdays within k days of each other given that there are no coincidental birthdays, we divide the number of distinct orderings that take this form by the total number of orderings. To find distinct permutations, we fix the first birthday and don't allow it to be permuted, as to eliminated "rotated" orderings. Rotated orderings are orderings that are sequentially the same but with different starting points.

Then, we group all of the 1's and the following $k-1$ 0's, and treat these as a unit. This ensures that our condition that no two birthdays are within the same interval of k days continues to be met throughout our permutations. We then permute all of the $0*$'s, of which there are $(365-1)-(n-1)-n(k-1)=365-kn$, and all of the $[1,\underbrace{0,0,\ldots,0}_{k-1}]$ units, of which there are $n-1$. The number of these groupings is:

$$\binom{(365-kn)+(n-1)}{n-1}=\binom{n-kn+364}{n-1}$$

By fixing the first birthday, the total number of potential orderings becomes $\binom{364}{n-1}$, so we can write our conditional probability as:

$$P(\overline{A_k}\mid\overline{A_1})=\frac{\binom{n-kn+364}{n-1}}{\binom{364}{n-1}}$$

Thus, the probability of having two birthdays within k days of each other in a group of size n is given by:

$$P(A_k) = 1 - \left(\frac{\binom{n-kn+364}{n-1}}{\binom{364}{n-1}} \cdot \frac{365!}{(365-n)!365^n} \right)$$

which simplifies to

$$P(A_k) = 1 - \left(\frac{(364-kn+n)!}{(365-kn)!365^{n-1}} \right).$$

This completes the solution to the Almost Birthday Problem. However, similar to the Basic Birthday Problem, this can be phrased in the more classical way:

> What is the smallest number of people required for a group so that the probability will be $\frac{1}{2}$ that two people in the group have a birthday within the same interval of k days?

This question is answered in a similar manner to the classical Basic Birthday Problem, by simply plugging values of n into the formula for a given interval k.

A collection of values for n given specific values of k is listed below:

k	n such that $p(n) = .5$
1	23
2	14
3	11
4	9
5	8
6	7
7	7
11	6

The Almost Birthday Problem was relatively straightforward, with an clear graphical interpretation of the problem providing a lot of insight. However, the next generalization we will go into is much more complicated.

4. Multiple Birthday Problem: Combinatorial Method

Another common generalization of the basic version of the Birthday Problem can be worded as thus:

> What is the probability that, among n people, at least m share some birthday?

This question is known as the Multiple Birthday Problem. After having gone through the solution of the basic version, the Multiple version may appear as the most natural generalization to make. Now, instead of just focusing on any 2 possible people sharing a birthday, we let this value m be anywhere from 2 to n. While there is more than one solution to the Multiple Birthday Problem, in this paper we present two effective solutions. The next section features Levin's Method, a probabilistic approach to this question. This section will discuss the Combinatorial Method, which takes on a more direct form of reasoning.

Since the Multiple Birthday Problem can account for many more possibilities than the basic version, it is important to take note of what kind of parameters we are dealing with. The probability that, among n people, at least m share some birthday can be written as a function of these parameters:

$f(n, m, c)$, where: n is the total number of people, m is the matching quota of interest, and c is the number of possible birthdays.

Note that while the Basic Birthday Problem is relatively easy to solve, we see that its solution addressed $f(n, 2, 365)$. With this in mind, the Basic Birthday Problem presents itself as a very specific question among a much larger class of problems.

Furthermore, we also see that setting $c = 365$ is consistent with our assumption that only that many birthdays exist in any given year, and for the purposes of any type of birthday problem this constant remains the same.

However, when this constant is relaxed and c is allowed to take on any positive integer, the number of possible questions may seem endless. At this point, instead of just birthdays, we are actually dealing with a very general multitude of questions called Cell Occupancy Problems.

Cell Occupancy Problems are processes where a number of items are distributed across a certain number of cells according to some probability distribution. For each cell, there is associated some kind of quota, where all the quotas may be the same or can be unique. The question is, then, finding the probability of whether any one, some, or all of the quotas get met in the process.

The Multiple Birthday Problem, specifically, is just a Cell Occupancy Problem where, again, c is fixed at 365 and the quotas for all the cells are set at a constant m. The n individuals are distributed over these cells according to their birthdays, where we stick with our assumption that all possible 365 birthdays are equally likely to occur. Our question, then, becomes whether *any* of the cells' quotas get met in the process.

Framing the Multiple Birthday Problem in terms of cells not only provides for a useful visualization, but it also serves as a reminder that the solutions presented in this paper can be used to address a wide variety of problems.

If we think about the Multiple Birthday Problem specifically, the birthdays of n people can be thought of as a process, or a collection of random variables:

$$B_i = \text{birthday of person } i, \qquad \text{where: } i = 1, 2, \ldots, n$$

$$B_1, B_2, \ldots, B_n \overset{iid}{\sim} \text{(discrete) Uniform}\{1, 2, \ldots, 365\}$$

where 1 represents January 1^{st}, 2 represents January 2^{nd}, ...

Thinking in these terms, our probability of interest, $f(n, m, c)$, can be thought of as the probability that at least m of the B_i's are equal. As in the basic version of the Birthday Problem, this will be approached by first finding the probability of the complement event:

$$f(n, m, c) = 1 - P(E)$$

where E is defined as the event that, for each birthday, the number of people with that birthday is at most $m - 1$.

In general, the way that $P(E)$ is computed is by summing the probabilities of all the ways n people can have birthdays, such that no birthday is shared by m or more people. In order to count all these possibilities, there needs to be a succinct method in describing all of them. One such way is to use the following definitions:

$$\begin{aligned} R_1 &= \text{number of nonrepeated } B_i\text{'s} \\ R_2 &= \text{number of pairs of equal } B_i\text{'s} \\ R_3 &= \text{number of triples of equal } B_i\text{'s} \\ &\vdots \end{aligned}$$

In the context of the Multiple Birthday Problem, R_1 can be thought of as the number of unique birthdays that exist among n people, R_2 the number of birthdays that are shared twice, etc. In terms of cells, R_1 is the number of cells that only have 1 item each, R_2 the number of cells that have 2 items each, etc.

Describing the breakdown of birthdays with this method is helpful, because for any collection of n people, their birthdays can be described by a specific arrangement:

$$\{n : r_1, r_2, \ldots, r_n\}$$

Recall that, ultimately, the complement probability $P(E)$ is found by considering all the possible ways n people can have birthdays such that no birthday is shared by m or more people. With the above description in mind, we can now find the following value:

$$P(n : r_1, r_2, \ldots, r_{m-1})$$

which is interpreted as: the probability that the birthdays of n people are broken down in a *specific arrangement*, such that no birthday is shared m or more times.

With the restriction that no birthday be shared m or more times, this value addresses a very high count of possibilities, and this is where the majority of the combinatorial argument comes in:

$$P(n : r_1, r_2, \ldots, r_{m-1}) = \frac{\frac{365!}{(365-r)!} \cdot \frac{n!}{\prod_{j=1}^{m-1}(r_j!)(j!)^{r_j}}}{365^n}$$

where $r = \sum_{i=1}^{m-1} r_i$.

First, notice that the formula appears as an extension of the formula in the Basic Birthday Problem. Recall that in the Basic Birthday Problem, since the matching quota was just 2, all the birthdays in the complement event had to be distinct, and all that was required was to permute the n different birthdays.

On the other hand, the matching quota may be higher than 2 in the Multiple Birthday Problem, and in the complement event E there may be overlap among people's birthdays. For this reason, this new r value is defined above, which keeps track of how many birthdays are actually being represented. r may also be thought of as the number of cells that actually have items in them. With all this in mind, the fraction $\frac{365!}{(365-r)!}$ immediately follows.

Second, when counting all the possible ways n people can have birthdays in a specific arrangement where no birthday is shared m or more times, the above permutations run into the problem of being *indistinguishable*. To account for this, the fraction $\frac{n!}{\prod_{j=1}^{m-1}(r_j!)(j!)^{r_j}}$ appears in the formula to discount any of the permutations that are indistinguishable and, in essence, to avoid over-counting. Thinking in terms of cells:

- $n!$ is the number of ways to order all of the n items,
- dividing by $\prod_{j=1}^{m-1} r_j!$ is done to remove redundancy among cells of the same size,
- and dividing by $\prod_{j=1}^{m-1}(j!)^{r_j}$ is done to remove redundancy among items within each cell.

When this entire procedure is done for all the j's from 1 to $m-1$, all indistinguishable permutations have been accounted for.

Finally, the entire numerator in our formula thus represents the number of *distinguishable* ways n people can have birthdays in a specific arrangement, such that no birthday is shared m or more times. When this value is divided by 365^n, which is just the total number of ways n people can have birthdays, we then arrive at the probability, $P(n : r_1, r_2, \ldots, r_{m-1})$.

Notice, however, that when $P(n : r_1, r_2, \ldots, r_{m-1})$ is computed, it is done for only *one* specific arrangement, $\{n : r_1, r_2, \ldots, r_{m-1})$. To compute $P(E)$, the above probability needs to be summed over all such possible arrangements:

$$\sum_{\{arrangements\}} P(n : r_1, r_2, \ldots, r_{m-1}) = P(E)$$

To be precise, these arrangements are exactly all integer partitions of n whose parts are at most $m-1$.

At this point, subtracting the complement probability from 1 results in our ultimate probability of interest:

$$1 - P(E) = f(n, m, c)$$

To calculate these probabilities, we put together a Python program (included in our submission folder as `partition.py`) that generates integer partitions with largest part $m-1$ as birthday arrangements and sums the probabilities of each arrangement occurring, which are calculated according to the method described above. This program is reliable up to about $m = 5$ or $m = 6$, but can get bogged down in generating the many necessary partitions for larger matching quotas and run very slowly. We were able to use it to verify the known results for smaller m, which will be presented at the end of the next section.

Although the Combinatorial Method counts every possible way the birthdays of n people can be broken down and provides an exact answer, carrying out the process can prove to be an extremely tedious task. Not only is computing $P(n : r_1, r_2, \ldots, r_{m-1})$ difficult in itself, but it must be done a very high number of times. In the next section, we describe a solution that is not exact in answer, but perhaps more elegant of a method.

5. Multiple Birthday Problem - Levin's Method

The previous combinatorial solution has the benefit of giving very exact results, because it is not reliant on any approximations. However, even restricting the size of the largest part to a very small value, the number of integer partitions required by that method's computations grows very quickly. For $m = 3$, the upper bookend for probability one-half, $n = 88$ gives rise to 45 legal arrangements, which is still borderline manageable, but for

$m = 4$, the upper bookend of $n = 187$ gives rise to 3008 partitions, at which point it becomes clear that calculating these probabilities without a computer program is out of the question.

This, together with the fact that this was intended to be a probability comps project, prompted us to investigate the probabilistic approaches to this problem; the one we will present here owes many of its details to Levin's treatment of Multinomial CDF's. As in the previous section, our goal is to find the probability that, among n people, at least m share some birthday. This time, we'll approach that question by way of a multinomial distribution, which extends the familiar Binomial distribution.

The Multinomial Distribution

In particular, we're given n items and c categories, where each category i is associated with a probability p_i of any given item falling into that category. Then if N_i gives the number of the n items that fall into the ith category, we have:

$$P(N_1 = n_1, \ldots, N_c = n_c) = \frac{n!}{n_1! \cdots n_c!} p_1^{n_1} \cdots p_c^{n_c}$$

Provided that $\sum_{i=1}^{c} n_i = n$. (And zero otherwise.)

To fit this to the multiple birthday problem, recall that we have:
n people in total (fixed)
$c = 365$ categories (birthdays)
$N_i = \#$ people out of n with birthday i, for $i = 1, 2, \ldots, 365$.
$p_i = \frac{1}{365}$ as the probability of each birthday occurring. (We assume a uniform distribution.)
m as the matching quota.

Then our Multinomial CDF takes the form:

$$P(N_1 = n_1, \ldots, N_{365} = n_{365}) = \frac{n!}{n_1! \cdots n_{365}!} p_1^{n_1} \cdots p_{365}^{n_{365}} = \frac{n!}{n_1! \cdots n_{365}!} \left(\frac{1}{365}\right)^n$$

Where we again require that $\sum_{i=1}^{365} n_i = n$.

To find the probability of an m-person match, we'll again use the complementary event E, that every N_i is smaller than m. Then:

$$P(E) = P(N_1 \leq m-1, N_2 \leq m-1, \ldots, N_{365} \leq m-1)$$

Which is a step in the right direction, but turns out to still be difficult to work with. To continue, we rely on a property of the multinomial distribution, which is more easily motivated in reverse.

The Poisson Connection

Suppose $X_1, \ldots, X_k$ are independent Poisson random variables with respective parameters λ_i, i.e. $P(X_i = x) = \frac{e^{-\lambda_i}\lambda_i^x}{x!}$ for $x = 0, 1, 2, \ldots$ Then:

$$(X_1, \ldots, X_k) \Big| \left(\sum_{i=1}^{k} X_i = n \right) \sim \text{Mult}(n, \tilde{\pi})$$

That is, conditional on that their sum is equal to n, the X_i's give the counts for a k-category multinomial distribution with parameters n and $\tilde{\pi}$, where $\pi_i = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_k}$.

We won't offer a proof here, but this makes some intuitive sense. Poisson variables are very frequently used to model counts, and though the infinite state space that they usually take would not seem to fit the multinomial framework, by conditioning on a fixed value for their sum we restrict them to give the counts out of n. Further the multinomial probability of an item falling into a given category, or here registering as a count for a given Poisson variable, is given by comparing its rate (λ) to the other Poisson variables.

Although this identity is more easily interpreted by starting with the Poisson variables, in the setting of the birthday problem we instead have the multinomial distribution, and so we will need to apply this distributional identity in reverse. Recall that N_i gives the count out of n birthdays of those falling on the ith day, and $(N_1, \ldots, N_{365}) \sim \text{Mult}(n, \tilde{\pi})$, Where $\pi_i = \frac{1}{365}$. Then by the previous identity, we can rewrite this distribution like so:

$$(X_1, \ldots, X_{365}) \Big| \left(\sum_{i=1}^{365} X_i = n \right) \sim \text{Mult}(n, \tilde{\pi})$$

Where the X_i's are now independent Poisson random variables. For the birthday problem, we have $\pi_i = \frac{1}{365}$, and to satisfy the Poisson setup we require $\pi_i = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_{365}}$, but these can be reconciled by setting $X_i \sim \text{Pois}(\frac{n}{365})$ for each i. (Levin mentions that any real number can be used in place of n, and it's true that the algebra will still check out at this level, but for the purposes of later computations it is most straightforward to choose $\lambda = \frac{n}{365}$.)

With this framework in place, we can now rewrite our target probability. Recall initially we had said that:

$$P(E) = P(N_1 \leq m-1, \ldots, N_{365} \leq m-1)$$

We can now instead write:

$$P(E) = P(X_1 \leq m-1, \ldots, X_{365} \leq m-1 \Big| \sum_{i=1}^{365} X_i = n)$$

Because, conditional on their sum, the independent Poisson random variables X_i are distributed identically to the N_i counts. Notice however that although it was necessary to condition on this sum in order to apply this property, in conditioning we lose the independence between the X_i's. To see this, notice that if it is given that $X_1+X_2+\ldots+X_{364}=n$, then X_{365} is equal to zero with probability one. To move forward with this argument, we next apply Bayes' Theorem, which gives:

$$P(E)=\frac{P(\sum_{i=1}^{365}X_i=n\big|X_i\leq m-1,\forall\, i)P(X_i\leq m-1,\forall\, i)}{P(\sum_{i=1}^{365}X_i=n)}$$

Putting Independence to Work

Now that the X_i variables are no longer conditioned on their sum, we can exploit some properties of independence. First, for the upper-right probability, we can instead write:

$$P(X_i\leq m-1,\forall\, i)=\prod_{i=1}^{365}P(X_i\leq m-1)$$

Due to the independence of the X_i's. Also applying that $X_i\sim \text{Pois}(\frac{n}{365})$ for all i (identical distribution), we can rewrite the product:

$$\prod_{i=1}^{365}P(X_i\leq m-1)=P(X_1\leq m-1)^{365}$$

The probability is now phrased in terms of a single Poisson random variable, and can be evaluated easily for any fixed m.

To continue decomposing $P(E)$, we require not just independence, but more specifically a convenient property of independent Poisson random variables. In particular, suppose $X_1,\ldots,X_k$ are independent Poisson random variables such that $X_i\sim\text{Pois}(\lambda_i)$. Then even if the λ parameters differ, we have that:

$$\sum_{i=1}^{k}X_i\sim\text{Pois}\left(\sum_{i=1}^{k}\lambda_i\right)$$

That is, the sum is also Poisson with parameter given by the sum of the parameters. This can be proved by way of probability generating functions, or through the combination of a conditioning argument and induction.

With this property in mind, recall that the denominator of $P(E)$ is given by $P\left(\sum_{i=1}^{365}X_i=n\right)$, and recall also that the X_i's are i.i.d. Poisson($\frac{n}{365}$). Applying the previous result, their sum is then Poisson with parameter $\lambda=\sum_{i=1}^{365}\frac{n}{365}=n$. Therefore:

$$P\left(\sum_{i=1}^{365}X_i=n\right)=\frac{e^{-n}n^n}{n!}$$

For some of our computations, n is sufficiently large that programs such as R will refuse to calculate n^n or $n!$ (for $m = 4$ we would like to be able to calculate these up to $n = 187$, which is not possible in R). In these cases, we can use Stirling's approximation, which states that $n!$ approaches $e^{-n}n^n\sqrt{2\pi n}$. The denominator of $P(E)$ is then conveniently reduced to $\frac{1}{\sqrt{2\pi n}}$. (If you'll recall the earlier statement that picking $\lambda = \frac{n}{365}$ for the X_i's resulted in the easiest computations, this is one of the places where that choice helps.)

One Last Piece

After the previous simplifications, we have:

$$P(E) = \frac{P(\sum_{i=1}^{365} X_i = n \big| X_i \leq m-1, \forall\, i) P(X_1 \leq m-1)^{365}}{\frac{1}{\sqrt{2\pi n}}}$$

Which leaves the upper-left probability for us to address. It turns out to be much more difficult to work with, but we can at least say this:

$$P\left(\sum_{i=1}^{365} X_i = n \big| X_i \leq m-1, \forall\, i\right) = P(W = n)$$

Where W corresponds to the sum of 365 independent <u>truncated</u> Poisson random variables, each with parameter $\frac{n}{365}$.

This leaves the question of what a truncated Poisson random variable is. The idea here is that instead of treating our sum as conditional on certain values for our Poisson variables, we can write it as an unconditional sum over truncated (range-restricted) Poisson variables; in effect we're just wrapping the conditioning statement into the variables in the sum. In particular, if we restrict each $X_i \sim \text{Pois}(\frac{n}{365})$ to the values $0, 1, \ldots, m-1$, then the corresponding truncated Poisson variable is $Y_i \sim \text{TPois}(\frac{n}{365})$, with pmf given by:

$$P(Y_i = x) = \frac{P(X_i = x)}{\sum_{k=0}^{m-1} P(X_i = k)} \text{ for } x = 0, 1, \ldots, m-1.$$

Which can be interpreted as the Poisson pmf, but normalized by a constant, the denominator sum, to ensure that the Y_i's follow a legitimate probability distribution.

Sadly, truncated poisson variables are hard to sum. Although independent Poisson random variables can be added very cleanly, the restricted range of the truncated variables makes the same computation much more complicated. However, our sum is over 365 random variables, so forces like the Central Limit Theorem start to come into play. In his paper, Levin states that a simple normal approximation is not accurate enough, and recommends using an Edgeworth correction, which is effectively a normal approximation that includes some corrective factors based on the moments of the random variables involved. Edgeworth approximations are somewhat outside of the scope of our project, so we took Levin's assertions on faith at this point; the calculations we performed can be replicated by referring to his work.

As a reminder before we present the results of this method, we currently have that the probability of seeing an m-person match is given by:

$$1 - P(E) \approx 1 - P(W = n)P(X_1 \leq m - 1)^{365}\sqrt{2\pi n}$$

Applying the Edgeworth approximation to $P(W = n)$, we found the following matching probabilities $(1 - P(E))$:

m	2 $(n = 23)$	3 $(n = 88)$	4 $(n = 187)$	5 $(n = 313)$
Levin	.5092	.5115	.5029	.5012
Combinatorial	.5073	.5111	.5027	.5011

Where the columns in this table give the values of the matching quota m, but also the smallest number n of people for which the probability of seeing a match for that value of m is above one half. Because the problem is often phrased in terms of finding these values of n, these probabilities are well-documented and allowed for the easiest verification of the accuracy of our calculations.

These results demonstrate that Levin's method for the multiple birthday problem gives results accurate to three and nearly four decimal places for large enough n. The slightly higher error in the $m = 2$ case (the simple birthday problem) is partially attributable to our use of Stirling's approximation for $n!$, which is more accurate for larger n. We were very pleased with these results, albeit somewhat surprised.

6. Conclusion

We began our process by exploring a simple generalization of the Birthday Problem, the Almost Birthday Problem, which was solved by way of a graphical visualization of orderings of birthdays. The Multiple Birthday problem was quite a bit more difficult to solve, and we approached it from two different methods of reasoning. The first one, the Combinatorial method, gave us an exact value for probabilities, but is unwieldy and not possible to do without automating the process. The second method, Levin's method, does use several approximations, but is doable mostly by hand and also gives us remarkably exact values for probabilities. A third method that we explored approached the problem by embedding the arrivals of the people (who give the birthdays) in a Poisson process, but we felt that even a light treatment of this method would require the introduction of a more involved framework than we could expect attendees to our comps talk to take in. We also felt that Levin's method, although it does eventually require approximations, is somewhat more of a direct probabilistic solution to the problem than the embedding argument, which also encouraged us to present it. As a final note, we assumed in all parts of this problem that all birthdays occur with equal probability (uniformly $\frac{1}{365}$). This is not quite true to life, as there is some evidence of small variations between months and between days of the week, but we saw some evidence that although theoretically our uniform assumption results in the lowest possible matching probabilities for fixed n and m, the empirical variability in birthday likelihood does not change the probabilities of matches occurring in any significant way. As a final note, if we do wish to expand these results past birthdays, then it is worth noting several of our approaches to the problem, including Levin's method, will accommodate different numbers of categories as well as nonuniform occurrence probabilities.

References

[1] P. Gorroochurn. *Classic Problems of Probability*, 2012, pages 240-246.

[2] B. Levin. A Representation for Multinomial Cumulative Distribution Functions. *The Annals of Statistics*, 1981, volume 9, no. 5, pages 1123-1126.

[3] E.H. Mckinney. Generalized Birthday Problem. *The American Mathematical Monthly*, 1966, volume 73, no. 4, pages 385-387.