Birthday Probability

Dahyun Hong, Shaambav Dave

November 10, 2023

1. Introduction

Problem 3 from Problem Set 1 in Orloff and Booth presents the following scenario:

> Ignoring leap days, the days of the year can be numbered 1 to 365. Assume that birthdays are equally likely to fall on any day of the year. Consider a group of *k* people, of which you are not a member. An element of the sample space Ω will be a sequence of n birthdays (one for each person).

Assigning each subsequent day of the year to natural numbers from 1 to 365 (i.e. January 1st = 1, January 2nd = 2, …. December 31st = 365), the set *S* can model the possible birthdays such that

$$S = \{1, 2, 3, \ldots, n\}$$

where n is the number of all possible birthdays in a given year. Ignoring leap years, we will assume $n = 365$. Since the sample space Ω is the set containing all sequences of S^k

$$S^k = \{b = (b_1, b_2, \ldots, b_k) : b_i \in S\}$$

We assume all birthdays are equally likely to occur (although we will see later that this is not the case), and there are 365^k sequences of k potential birthdays. The uniform probability measure assigned to each sequence of birthdays is

$$\pi(b) = \frac{1}{365^k}$$

In particular, we are interested in finding out the probability that given a collection of k people, what is the probability of at least m of them sharing the

same birthday? We will formulate a mathematical model to investigate the probability of shared probabilities within a group of people, beginning with the base case $m = 2$ before expanding to explore the probabilities for different values of m and k ($m = 3$, $k = 3$, $k = 4$). We will also simulate the event of m shared birthdays using the colMatches function in R. Finally, we will look at how our expected uniform distribution of birthdays does not hold true in real life (some birthdays are more common than others), and how this skewed distribution increases the probability of matching birthdays for a fixed k.

2. Model

As established in the introduction, for the purposes of the model we assume that all birthdays to be equally likely to occur such that

$$\pi(S^k) = \frac{1}{365^k}$$

Let us establish the set D_k, representing the set of k-tuples with distinct birthdays:

$$D_k = \{b \in S^k : b_i \neq b_j \ , i \neq j\}$$

Where, since k and n are distinct elements with no repetitions allowed, the number of elements in D_k

$$\#D_k = P(n,k) = \frac{n!}{(n-k)!}$$

which is equivalent to the count of different ways of selecting k distinct birthdays from a total of n distinct possible dates. If $k > n$, $\#D_k = \phi$, since at least one birthday must overlap.

Let us also establish $M_l(k)$, the set of k-tuples where exactly ℓ individuals share the same birthday within the group of k people -- it contains sequences where the most frequently occurring birthday within the group of k people happens ℓ times.

$$M_l(k) = \{b \in S^k : max_{j=1,\dots,k} \#\{i \in \{1,\dots,k\} : b_i = b_j\} = l\}$$

The union of sets M_l for l ranging from m to k *r*epresents the combined set of sequences where the maximum count of any specific birthday within the sequence falls within the range from m to k.

$$\bigcup_{l=m}^{k} M_l$$

This union captures all sequences where there are at least m shared birthdays within the group of k people. To find the probability of at least m shared birthdays within a group of k people, we find the product measure of this set such that:

$$\pi\left(\bigcup_{l=m}^{k} M_l\right) = \frac{\#\left(\bigcup_{l=m}^{k} M_l\right)}{\#\Omega} = \frac{1}{365^k}\sum_{l=m}^{k} \#M_l$$

We must also set some bounds for m and k before going further into our analysis. The probability of at least one person sharing the same birthday (m = 1) is 1, because every person at least shares a birthday with themselves. We can also reasonably deduce that most k people can share the same birthday--that is, every single person in the collection has the same birthday. The probability again approaches 1 when k becomes too big--if, for example, m = 2, and k > 365, then there must be at least two people sharing the same birthday, since at most only 365 people can have unique birthdays in a given year. Thus, a more practical approach to the problem would be to investigate the cases where $2 \leq m \leq k \leq (m-1)n$.

3. Case 1: $m = 2$

We want to find the probability that at least 2 among k people share the same birthday, or

$$\bigcup_{l=2}^{k} M_l$$

Another way to approach this problem is by assuming that at most one birthday is shared, which is equivalent to the statement that every birthday is distinct. The complement to this event would be that at least two people share the same birthday. Thus we know that

$$\bigcup_{l=2}^{k} M_l = S_k \backslash M_1$$

where $M_1 = D$. We've already established $\#D_k = P(n,k)$ while k is at most n. By the property of the complement of a set $M_2 \subset S_k$,

$$\sum_{l=m}^{k} \#M_l = n^k - \frac{n!}{(n-k)!}$$

Substituting this into the formula for the product measure on $\bigcup_{l=m}^{k} M_l$ that we found earlier, we find that

$$p(k;2,n) = \frac{1}{n^k}\left(n^k - \frac{n!}{(n-k)!}\right) = 1 - \frac{n!}{n^k(n-k)!}$$

we also know that $n = 365$, therefore for $k \leq 365$,

$$p(k;2,n) = 1 - \frac{365!}{365^k(365-k)!}$$

When k exceeds n, $p(k;2,n) = 1$ since there can only be n unique birthdays, and so there must be at least two people whose birthdays overlap.

4. Case 2: $m = 3$

a. $k = 3$

We want to find the probability that all 3 people share the same birthday, or

$$\bigcup_{l=3}^{k} M_l, \; k = 3$$

In the case where $m = 2$, we did not have to directly compute the value of $\sum_{l=2}^{k} \#M_l$ because we found the cardinality of the complement to M_2. We could try to find the probability that two or less people share the same birthdays and use the property of set complements like we did in case 1, which would be tedious. However, we know that $\sum_{l=3}^{3} \#M_l = 365$, since all 3 people share the same birthday and there are 365 possible days that the shared birthday could fall on.

$$p(3; 3, n) = \frac{n}{n^k} = \frac{1}{n^{k-1}} = \frac{1}{365^2} \approx 7.506 \times 10^{-6}$$

We can also arrive at this solution by breaking apart the problem more intuitively--person A could have a birthday on any day of the year, and the probability of them having a birthday that falls between 1 and 365 is 1. We have already established that the sample space Ω admits a uniform product measure, and thus the probability of persons B and C being born on that particular day would be $\frac{1}{n} = \frac{1}{365}$. By the Bayes Theorem, assuming each persons' birthday is independent of the others, we find

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C) = \frac{1}{365^2}$$

b. $k = 4$

Things become slightly more challenging when $k = 4$. Again, we cannot use the same formula derived from Case 1. Instead, we can split this into two parts: there are exactly 3 people who share the same birthday, and there are exactly 4 people who share the same birthday.

To calculate the probability that there are exactly 3 shared birthdays, we must manually calculate the sum of the count of *k*-tuples with exactly 3

shared birthdays. There are $P(4,3) = 4$ ways that 3 out of 4 people can share the same birthday. There are 365 potential days that the shared birthday can fall on, and 364 potential days for the distinct birthday.

$$\frac{P(4,3)\times 365\times 364}{365^4} = 2.994 \times 10^{-5}$$

To calculate the probability that there are exactly 4 shared birthdays, the process is similar to Case 2.a, where we found 3 shared birthdays among 3 people. We know that $\sum_{l=4}^{4} \#M_l = 365$, since there are $P(4,4) = 1$ way that all 4 people can share the same birthday, and there are 365 potential dates that the same birthday can fall on.

$$p(4; 4, n) = \frac{1}{n^{k-1}} = \frac{1}{365^3} \approx 2.056 \times 10^{-8}$$

Summing the two cases, we find that

$$p(4; 3, n) \approx 2.996 \times 10^{-5}$$

5. Other Things to Consider

The following scenarios are presented in Orloff and Booth:

a. Event A: "someone in the group shares your birthday"

Note that you are not a part of the group of *k* people.

(Solution) Taking a similar approach to the case that *m = 2*, if we calculate the probability of event A^c: no one from the group shares the same birthday as you and subtract that from the probability of the entire sample space Ω, which is always 1, we can find $P(A)$ by the property of set complements. $\#A^c = 364^k$, since for each *k* persons there are 364 potential birthdays that do not overlap with your birthday, and $\#\Omega = 365^k$.

$$P(A) = 1 - P(A^c) = 1 - \frac{364^k}{365^k}$$

b. Find the minimum number of people *k* such that p(k; m, n) meets a probability threshold *q*

For example, what is the smallest *k* such that the probability of at least two people sharing the same birthday is greater than 0.50?

(Solution) Using the formula from Case 1, we can rearrange the variables to find that for *k > 22,* the probability of at least two shared birthdays exceeds 0.50.

$$1 - \frac{n!}{n^k(n-k)!} \geq q \text{ or } n^k(n-k)! \geq \frac{n!}{1-q}$$

We can also use this formula on the event from 5.a to find the smallest *k* such that

$$P(A) = 1 - \frac{364^k}{365^k} = 0.5$$

(Solution) Rearranging the variables, we get

$$\frac{364^k}{365^k} = 0.5$$

which we can further simplify down to

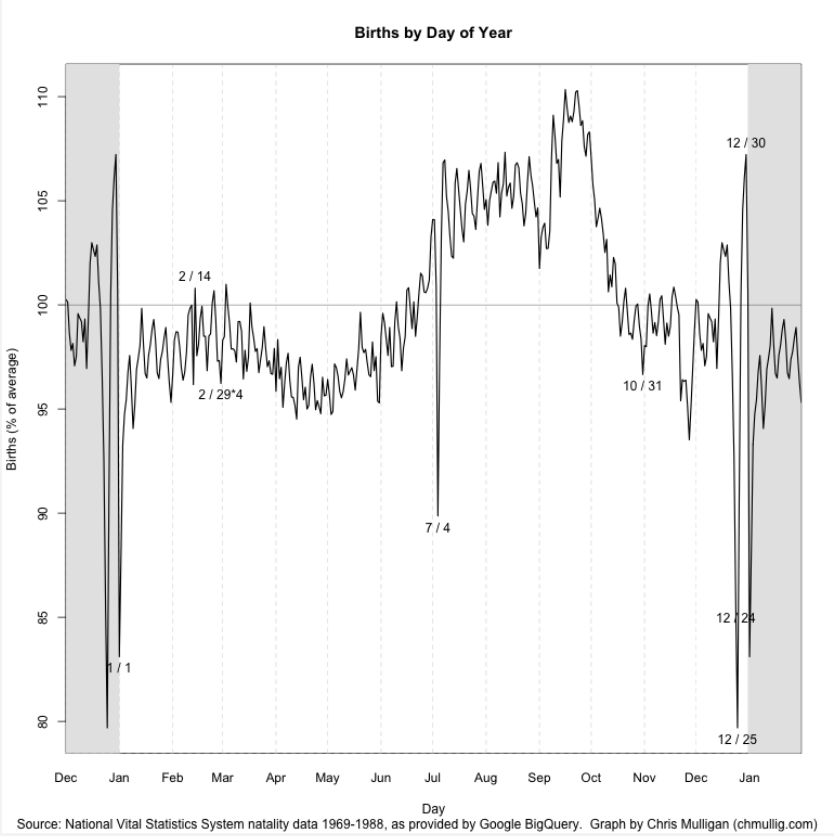
$$k \times ln(\frac{364}{365}) = ln(0.5)$$

$$k \approx 252.65$$

Thus, the smallest *k* such that $P(A) > 0.5$ is 253 people. Repeating trials using R simulation yields similar results. Notice that 253 is greater than 365/2 -- this is because the birthdays are not guaranteed to be unique. 365/2 people may have overlapping birthdays, thus decreasing the probability of matching your birthday.

6. Real-Life Application

In reality, some birthdays tend to occur more frequently than others, as displayed by the graph from Chris Mulligan below. Note that this data comes from 2012, over a decade ago, and thus may not be the most accurate--it still presents an interesting aspect of the birthday problem to take into consideration.



The birth rate on average increases during the first half of the year, reaching its peak around September, before declining. There are a few interesting spikes as well as drops on specific days as well--notably centered around global holidays like Christmas. Since some birthdays occur at a higher frequency than others (i.e. December 30th versus January 1st), for a fixed number of k people, a matched outcome is more probable than using the equal-probability model that we've adopted for our analysis.

7. Appendix: colMatches

The colMatches function, which is an 18.05 function, within the context of birthday probability, counts the number of shared birthdays or matches within a group of individuals

Simulating at least 2 people sharing the same birthday (*m = 2*)

```
# Setting up parameters
source("colMatches.r")
days = 365  # Total number of days in a year
people = 25  # Number of people in each trial
trials = 10000  # Number of simulation trials
sizematch = 2  # Desired size of the match (at least two people
sharing a birthday)

year = 1:days  # Days in a year

# Generate random birthdays for all trials
y = sample(year, people * trials, replace = TRUE)
trials = matrix(y, row = people, col = trials)

# Use colMatches function to count matches of size sizematch
within each trial
matches = colMatches(trials, sizematch)

# Calculate the probability of having at least sizematch people
sharing a birthday within each trial
prob_match = mean(matches)
```

To simulate at least 3 people sharing the same birthday (*m = 3*), we just update sizematch to 3 for the same code.