A binomial distribution approximation to the generalized birthday problem

Wei Ruen Leong

December 23, 2018

Abstract

The generalized birthday problem concerns a set of events that has to be found by solving a subset sum problem from a multinomial distribution. The subset sum problem is hard to solve when the magnitude of the parameters increases. I propose a binomial distribution approximation for the generalized birthday problem, and I supply some numerical examples to compare the accuracy of the approximation with the true probabilities.

1 Introduction

Consider the generalized birthday problem: What is the probability of at least a minimum of k people sharing a common birthday in a calendar year of T days in a room of n people, where k, n, and T are positive integers? Assume that everyone must have one and only one birthday, and the birthday must falls in one of the T days. Furthermore, assume that each birthday is equally likely; birthdays are uniformly distributed, with each birthday having a probability of $1/T$ of occurring.

The classical birthday problem considers the case of $k = 2$ and $T = 365$ (the leap day, February 29 is excluded for simplicity), i.e. what is the probability of at least a minimum of two people sharing a common birthday in a room of n people? For $n = 23$, the probability can be computed to be approximately 0.50.

The sample space in the generalized birthday problem is identical to rolling a (fair) T-sided die n number of times, it can be characterized by a multinomial distribution. To elaborate, let X_i be random

variable representing the number of times the ith day is observed as a birthday for $i \in \{1, 2, \ldots, T\}$, then X_i follows a multinomial distribution with parameters, n and $p_j = 1/T$ for $j \in \{1, 2, \ldots, T\}$, and with support $x_i \in \{0, 1, \ldots, n\}$ with the restriction that $\sum_{i=1}^{T} x_i = n$. The set of events that is of interest for the generalized birthday problem is the set of events with permutations of $x_i \in \{0, 1, \ldots, k-1\}$, such that $\sum_{i=1}^{T} x_i = n$. To find all the possible permutations, one has to solve a subset sum problem, that is a special case of the knapsack problem.

2 The classical birthday problem formula

It turns out for the classical birthday problem ($k = 2$, $T = 365$), an analytical formula can be derived easily to compute the desired probability. First, consider the case of $n = 366$, then the probability is always 1 by the pigeonhole principle. To illustrate this principle, take the worst case scenario, where the first 365 people have distinctive birthdays. When it comes to the 366th person, the person birthday will fall in one of the 365 days, thereby sharing the same birthday with one of the first 365 people. The same pigeonhole principle applies for $n \geq 366$.

Consider the case of $k = 2$ and $n < 366$ instead. I introduce the following notations for my computations:

$$\begin{aligned} P(M) &: \text{the desired probability to be computed, } P(M) \\ &\quad \text{means the probability of a match in short.} \\ P(M') &: \ P(M') \text{ means the probability of no match in short,} \\ &\quad \text{and } P(M) = 1 - P(M'). \\ \binom{n}{r} &: \ \text{the } n \text{ choose } r \text{ formula, given as } \frac{n!}{r!(n-r)!}. \end{aligned}$$

It is easier to find the formula for $P(M')$ than it is for $P(M)$. It can be given as follows,

$$P(M') = \frac{\binom{T}{1}\binom{T-1}{1}\binom{T-2}{1}\cdots\binom{T-n+1}{1}}{T^n}. \tag{1}$$

Equation (1) is the probability of n people having distinctive birthdays in a calendar year with T days. The denominator in (1) represents the total number of all possible birthdays in a room of n people. The numerator represents the total number of all possible distinctive birthdays.

To elaborate, the first term is $\binom{T}{1}$, it means the first person can his or her one birthday from T possible days. In other words, $\binom{T}{1}$ represents the total number of possible birthdays the first person can have from T calendar days. Moving on to the second term, $\binom{T-1}{1}$, it represents the total number of possible birthdays the second person can have from the remaining $T-1$ calendar days. From there on, a similar rolling fashion continues until the term $\binom{T-n+1}{1}$ in the numerator. Equation (1) can be alternatively expressed as follows,

$$\begin{aligned} P(M') &= \frac{T \times (T-1) \times (T-2) \times \ldots \times (T-k+1)}{T^n} \\ &= \frac{T!}{(T-k)!T^n}. \end{aligned}$$

To sum up, the probability of matching at least two ($k = 2$) or more people who share the same birthday in a room of n people can be computed with the following formula,

$$P(M) = \begin{cases} 1 - \frac{T!}{(T-k)!T^n}, & 0 < n < T. \\ 1, & n \geq T+1. \end{cases} \tag{2}$$

Now, what if $k > 2$? The formula for $P(M)$ is still harder to find compared to $P(M')$, however, $P(M')$ becomes much more complicated to compute compared to the one in (1) when $k = 2$. Take $k = 3$ as an example, then $P(M')$ is not just the probability of n distinctive birthdays. There are other events to consider when computing $P(M')$. I will state some of the events here, and the events I state are by all no means exhaustive. The events include a pair of people sharing the same birthday, two pairs of people sharing the same birthday, and so forth until $[n/2]$ pairs, where [.] denotes the integer part of a number. For the interested readers, the analytical formula for $P(M')$ for $k = 3$ is provided in DasGupta (2005) using multinomial probabilities. It will be useful to have an approximation formula to compute $P(M')$ as the true formula becomes increasingly harder to compute analytically when k increases from 2. In the next section, I will introduce an approximation to the $P(M')$ formula for any positive integer k.

3 A binomial distribution approximation

Take the classical birthday problem, where $k = 2$ in the previous section as an example. I consider the number of possible pairs out of n people, that is the number $\binom{n}{2}$. The probability of a pair of people sharing their birthdays in a specific day is $\frac{1}{T} \times \frac{1}{T} = \frac{1}{T^2}$. There are T days in a calendar year, therefore the probability of a pair sharing birthdays in any day is $T \times \frac{1}{T^2} = \frac{1}{T}$. The approximation for $P(M')$

is then given by binom(0, $\binom{n}{2}$, $\frac{1}{T}$), where binom(x, m, p) is the binomial probability mass function with parameters x, m, and p. For $x = 0$, binom(0, m, p) can be simplified to just $(1-p)^m$. The formula binom(0, $\binom{n}{2}$, $\frac{1}{T}$) computes the probability of having a zero match in common birthdays for $\binom{n}{2}$ number of possible pairs. It is an approximation to the exact formula represented in (1) because by confining the probabilistic analysis to pairs, it excludes the probability of more than two people sharing the same birthday. For example, it does not account for the probability of a triplet of people sharing the same birthday. After all, I am approximating a multinomial distribution, that is the true distribution with a binomial distribution.

The general approximation formula for any positive integers k, n and T, with $k \leq n$, is given as follows,

$$P(M)^{\text{approx}} = \begin{cases} 1 - \frac{1}{T^{k-1}}^{\binom{n}{k}}, & 0 < n < (k-1)T + 1. \\ 1, & n \geq (k-1)T + 1. \end{cases} \tag{3}$$

I provide some numerical examples of how accurate the approximation is for $k = 2$ and $k = 3$ in Table 1 and Table 2 respectively. By inspecting the figures on the tables, I suspect the accuracy of the approximation worsens as k increases. Another noteworthy point is that the approximation is close to the true, at small values of n, then as n increases, the distance becomes further apart from the true, with the distance peaking at some n's, and finally the distance becomes closer again after the peak point(s).

n	$Approx$	$True$	$Approx - True$
5	0.027	0.027	0.000
10	0.116	0.117	-0.001
23	0.500	0.507	-0.007
30	0.697	0.706	-0.009
40	0.882	0.891	-0.009
50	0.965	0.970	-0.005
60	0.992	0.994	-0.002

Table 1: The second and third columns of the table report the approximate and true probabilities of the generalized birthday problem with parameters $k = 2$ and $T = 365$ for various values of n. The last column reports the difference between the approximate and true probabilities. The approximate probabilities are computed from (3), whereas the true probabilities are computed from (2).

n	$Approx$	$True$	$Approx - True$
23	0.013	0.013	0.000
40	0.072	0.067	0.005
60	0.227	0.207	0.020
75	0.398	0.361	0.037
87	0.549	0.499	0.050
88	0.561	0.511	0.050
145	0.976	0.952	0.024

Table 2: The second and third columns of the table report the approximate and true probabilities of the generalized birthday problem with parameters $k = 3$ and $T = 365$ for various values of n. The last column reports the difference between the approximate and true probabilities. The approximate probabilities are computed from (3), whereas the true probabilities are extracted from DasGupta (2005).

References

DasGupta, A. (2005), 'The matching, birthday and the strong birthday problem: a contemporary review', *Journal of Statistical Planning and Inference* **130**, 377–389.