Birthdate Identifier Deduplication

Daniel Alvarez, CBT Data Assurance

September 6, 2019

Birthdate Identifier Deduplication

This document presents the mathematical proof to determine likelihood of two groupings of birthdate identifiers being the same.

Presentation of the problem

A concern raised in the provision of humanitarian assistance in the context of a country with multiple humanitarian programs has been the possibility of beneficiaries receiving more than their entitled benefits across multiple programs. An example would be a beneficiary family successfully registering across multiple programs, using different forms of identification, in order to receive more than their entitled benefits. In this scenario, humanitarian providers want to be able to identify such beneficiaries and perform an adjudication of identity. The problem arises that identifying beneficiaries may be complicated by the use of different forms of identifying documentation and different identifying information set across different program registration processes.

Examining the dataset template fields across multiple programs, it was found that birthdate would likely be an mandatory identifier captured across multiple program registrations and, assuming that transcription error is low in the registration process, could be used to uniquely identify beneficiary households. Taking the set of birthdates for all individual households together as a single identifier, this identifier could serve to uniquely identify beneficiary households since the likelihood of a given beneficiary household having the same set of birthdates as another household is extremely low.

What follows is the mathematical proof of the likelihoods of duplication in household birthdate sets. The question this answers is: how likely is it that two households have the same set of birthdates?

Proof of mathematical concept

The birthdate set deduplication problem relies on a probabilistic model to reduce the complexity of finding a hash function.

Let:

- k = individuals in a household (i.e. household size)
- n = possible birthdates
- M = households
- Z = possible sets of birthdates
- $P(M; Z)$ = probability that at least one set of birthdates appears more than once out of M households
- $M(P^*; Z)$ = the smallest number of households M such that the probability that at least one set of birthdates appears more than once is at least some value, P^*

We can determine Z using a combination assuming n possible birthdates and k individuals in a household as follows:

$Z = C(n, k) = \frac{n!}{(n-k)!k!}$

Using a Taylor series expansion, we can also define the probability that at least one set of birthdates appears more than once out of M households, $P(M; Z)$:

$P(M;Z) = 1 - e^{\frac{-M(M-1)}{2Z}}$

By arithmetic manipulation, we can determine the smallest number of households M such that the probability that at least one set of birthdates appears more than once is at least some value, P^*:

$M(P^*;Z) = \sqrt{2Z * ln(\frac{1}{1-p})}$

Worked example

Lets assume:

- 365 days in a year (no leap years)
- 80 maximum possible years of age
- possible birthdates for a given individual, n = 365 * 80 = 29200.
- there are k=4 individuals in a household (i.e. household size of 4)
- there are M=10000 possible households

Then, $Z = C(29200, 4) = \frac{29200!}{(29200-4)!4!} = 302852331854527$

Therefore, the probability of one household out of M possible households sharing the same set of birthdates with any other household:

$P(M;Z) = 1 - e^{\frac{-M(M-1)}{2Z}} = 1 - e^{\frac{-10000(10000-1)}{2*302852331854527}} \approx 0$

The smallest number of households M such that the probability that at least one household shares the same set of birthdates with any other household is 1%:

$M(P^{1\%};Z) = \sqrt{2Z * ln(\frac{1}{1-p})} = \sqrt{2(302852331854527) * ln(\frac{1}{1-.01})} = 24672931$

Probability table examples

For further exposition, Table 1 below shows the probabilities of two groups sharing the same set of birthdates under under different parameter values. Here, we assume birth year is part of birthdate and the maximum age is 80.

Table 1: Probabilities accounting for birth year

ï..HouseholdSize	Total_Combinations	Households	Probabilities
1	29200	10,000	100.00%
2	426,305,400	10,000	11.07%
3	4,149,088,356,400	10,000	0.00%
4	30,285,233,185,452,700	10,000	0.00%
5	176,841,533,616,495,000,000	10,000	0.00%
6	860,481,428,988,930,000,000,000	10,000	0.00%

Additionally. Table 2 below shows the probabilities assuming birth year is not part of birthdate. That is, birthdates for a given individual can only take on 365 possible values (excluding the possibility of leap year birthdates).

Table 2: Probabilities not accounting for birth year

ï..HouseholdSize	Total_Combinations	Households	Probabilities
1	365	10,000	100.00%
2	66,430	10,000	100.00%
3	8,038,030	10,000	99.80%
4	727,441,715	10,000	6.64%
5	52,521,291,823	10,000	0.10%
6	3,151,277,509,380	10,000	0.00%

Assumptions

This proof relies on several underlying assumptions:

1. Variations in the distribution, such as leap years or twins (i.e. several household members with the same birthdate) are disregarded. Assumes 365 days in year.
2. Equal likelihood for all 365 possible birthdates for a given individual. In reality, birthdate distributions are not uniform since birthdates are not equally likely, but this irregularity has little affect on the proof. In fact, using a uniform distribution of birthdates is the worst case.
3. Order of birthdates within a group's set of birthdates does not matter.

References

Morton Abramson and W. O. J. Moser. "More Birthday Surprises". *The American Mathematical Monthly* Vol. 77, No. 8 (Oct., 1970), pp. 856-858.

William Knight and D. M. Bloom "A Birthday Problem". *The American Mathematical Monthly* Vol. 80, No. 10 (Dec., 1973), pp. 1141-1142.