The Birthday Problem when Birthdays are Dependent

Terry R. McConnell
Syracuse University

September 29, 2022

Abstract

We show that when birthdays in a group of people are suitably positively correlated, it is more likely that there will be a birthday in common than when birthdays are independent.

1 Introduction

According to the solution of the classical *Birthday Problem*, in any random group of 23 or more people the odds are better than even that some pair of people will share a common birthday.

To make a theorem out of this, we must be more precise about assumptions. Typically, it is assumed that birthdays of individuals are independent of each other and equally likely to occur on any of the 365 days of the year. The latter assumption of uniform distribution, however, is quite unnecessary for the desired conclusion. As has been noted by several authors, non-uniform distribution of birthdays only makes it more likely that some birthday will be represented more than once in a group. See, for example, [2], where this result is derived from an inequality about certain multi-linear averages.

What about the assumption of independence? If some degree of positive correlation held among individual's birthdays, then it seems likely that there would be an enhanced probability of having a common birthday in a group of given size. The birthdays of a group of 23 strangers who happen to attend one of many breakout sessions at a typical large convention might reasonably be modelled as independent. On the other hand, if the convention were the annual meeting of the Society of Identical and Fraternal Twins (SIFT)[1], then their birthdays would be positively correlated, and as soon as both members of a set of twins decide to attend the same session, concordance of birthdays is guaranteed.

An immediate problem is how to define a suitable notion of positive correlation? For a pair X and Y of real-valued random variables, positive correlation means that we have $E(XY) \geq E(X)E(Y)$. While we can certainly assign numbers to days of the year, this seems somewhat unnatural and arbitrary. Moreover, it is likely that one would need to assume a notion of positive correlation that applies not only to pairs of random variables, but to larger groupings as well.

An obvious candidate for a suitable notion of positive correlation is that of *association* of random variables introduced in [1]. Association imposes a strong form of positive correlation among a set of random variables and each of its subsets, and it is most naturally formulated for random variables taking values in an abstract totally ordered set. It would be quite reasonable to conjecture that concordance of birthdays is at least as likely among a group of associated birthdays as among a group of independent birthdays having the same marginal distributions.

Perhaps surprisingly, it turns out that association alone is not sufficient for this conclusion, even for pairs of birthdays. A more uniform kind of association appears to be required. We formulate and prove a theorem to this effect in section 3 of this paper, after having reviewed some relevant facts about association in the following section. Section 4 is devoted to illustrative examples. The last section discusses some open problems.

[1] A fictional organization, as far as I know

In order to avoid technicalities, we assume throughout that all random variables take values in a given *finite* totally ordered set S.

2 Associated Random Variables

Let $X_1, X_2, \ldots, X_N$ be a collection of S-valued random variables defined on a common probability space. The collection is said to be *associated* if, given any pair of functions $F, G : S^N \to \mathbb{R}$ which are non-decreasing in each variable, we have

$$EF(X_1, \ldots, X_N)G(X_1, \ldots, X_N) \geq EF(X_1, \ldots, X_N)EG(X_1, \ldots, X_N).$$

This definition was introduced in [1], and that reference should be consulted for the facts cited in this section.

It is true, but not trivial, that a collection consisting of a single random variable is associated. This follows from a classical inequality due to Chebyshev. By successive conditioning, it then follows that families consisting of independent random variables are associated. Since $f(x) = F(x, x)$ and $g(x) = G(x, x)$ are non-decreasing whenever F and G are non-decreasing in each variable, it follows that a family X, X, where X is any random variable, is associated. The variables in each of these cases are positively correlated or, at worst, uncorrelated; indeed, association in general may be viewed as a strong form of positive correlation amongst the variables.

An attempt to verify that a given collection is associated on the basis of the definition is unlikely to succeed except in the very simplest examples. The task is eased somewhat by the observation that it suffices to consider Boolean functions F and G, but even if the random variables are also Boolean, the number of cases that must be checked grows very rapidly with N. For example, there are 20 distinct non-decreasing Boolean functions of three Boolean variables, and so 400 different pairs of such functions must be considered to fully check the definition. (The number of monotone Boolean functions is counted by a sequence indexed by N known as the *Dedekind sequence.*)

Let T be another totally ordered set and $F_i : S^N \to T, i = 1, 2, \ldots M$, be functions that are each non-decreasing in each variable. Then if $X_1, X_2, \ldots, X_N$ are associated with values in S, the variables $Y_i = F_i(X_1, X_2, \ldots, X_N), i = 1, 2, \ldots, M$ are associated with values in T. Thus, an effective way to produce examples is to start with an independent collection and apply functions that are non-decreasing in each variable.

Let $b_1, b_2, \ldots, b_N$ be elements of S. Let $F(x_1)$ be the function that equals 1 if $x_1 \geq b_1$ and zero otherwise, and let $G(x_2, \ldots, x_N)$ be the function that equals 1 if all of the x_i are greater than or equal to the corresponding b_i and zero otherwise. Then both F and G are non-decreasing in each variable, hence

$$EF(X_1)G(X_2, \ldots, X_N) \geq EF(X_1)EG(X_2, \ldots, X_N).$$

Iterating this, we find that

$$P(X_1 \geq b_1, X_2 \geq b_2, \ldots, X_N \geq b_N) \geq P(X_1 \geq b_1)P(X_2 \geq b_2) \ldots, P(X_N \geq b_N).$$

One should beware that some natural operations do not preserve association. For example, mixtures of associated sequences are not necessarily associated, and random samples taken from associated collections are not necessarily associated, even if the sampling protocol is independent of the collection.

3 Main Results

Given a collection $X_1, \ldots, X_N$ of random variables defined on a common probability space, we denote by $\mathcal{F}_{\hat{i}}$ the sigma field generated by all variables except X_i. We say the collection is *affiliated* if and only if, for any subset A of S and index i, we have

$$P(X_i \notin A | \mathcal{F}_{\hat{i}}) \leq P(X_i \notin A) \text{ on the event } \bigcap_{j \neq i} \{X_j \in A\}, \tag{3.1}$$

and the same holds for any subcollection.

Stated informally, the variables “want” to belong to sets the other variables belong to.

Clearly, a single random variable is affiliated. The following properties of affiliation are useful in producing examples:

a. A pair X, X is affiliated, where X is any random variable.

b. A non-random collection is affiliated.

c. A subcollection of an affiliated collection is affiliated.

d. Independent affiliated collections can be merged, and the resulting collection is affiliated.

The first 3 statements are obvious. To show (d), let $X_1, X_2, \dots, X_N$ be affiliated, and let Y be independent of this collection. Let A be a subset of S. Then

$$P(X_i \in A | \mathcal{F}_{\hat{i}} \vee \sigma(Y)) = P(X_i \in A | \mathcal{F}_{\hat{i}}) \geq P(X_i \in A)$$

on the event $\bigcap_{j \neq i} \{X_j \in A\}$ and subsets. Also, we have $P(Y \in A | \sigma(X_1, X_2, \dots, X_N)) = P(Y \in A)$. Similar reasoning applies if Y is a vector with affiliated components.

The birthdays of people attending the SIFT convention can be modeled as an affiliated sequence: if the equal pairs of birthdays of twins attending the conference are independent of each other, then the collection of all birthdays at the conference is affiliated by (a) and (d).

Theorem 3.1. *Let $X_1, X_2, \dots, X_N$ be affiliated random variables and $Y_1, Y_2, \dots, Y_N$ be independent random variables such that X_i has the same distribution as Y_i for $i = 1, 2, \dots, N$. Then*

$$P(X_1, X_2, \dots, X_N \textit{ are distinct }) \leq P(Y_1, Y_2, \dots, Y_N \textit{ are distinct }).$$

Let A be a given subset of S having cardinality $N - 1$. Then using (3.1) we have

$$P(X_N \notin A | \{X_1, X_2, \dots X_{N-1}\} = A) P(\{X_1, X_2, \dots, X_{N-1}\} = A) \leq P(Y_N \notin A) P(\{X_1, X_2, \dots, X_{N-1}\} = A).$$

Summing these inequalities over all choices of A produces

$$P(X_1, X_2, \dots, X_N \text{ are distinct }) \leq P(X_1, X_2, \dots, X_{N-2}, Y_N, X_{N-1} \text{ are distinct }).$$

By (c) and (d), $X_1, X_2, \dots, X_{N-1}, Y_N$ is affiliated. Thus, we may repeat the argument with X_{N-1} playing the role of X_N, etc.

If we also assume the variables in the affiliated collection are identically distributed, then we may combine the last theorem with the results of [2] to obtain an extension of the Birthday Problem.

Theorem 3.2. *Let $X_1, X_2, \dots, X_N$ be affiliated and identically distributed random variables with values in a set of cardinality K, with $K \geq N$. Then*

$$P(X_1, X_2, \dots, X_N \textit{ are not distinct }) \geq 1 - \frac{(K-1)(K-2)\dots(K-N+1)}{K^{N-1}}.$$

These theorems do not apply directly to the SIFT conference, since the hypotheses are not satisfied: the people attending a given session are a random sample of the people at the conference, and we have no result that says a random sample from an affiliated collection is affiliated. On the other hand, the conclusions of both theorems do apply, since we can condition on the sample chosen on both sides.

4 Examples

We provide an example of a pair of random variables that is associated but such that the result of Theorem 3.1 does not hold. This collection is therefore associated but not affiliated.

The joint distribution of (X, Y) is given by the following table, where for now we only assume p is chosen so that $0 < p < 1$:

probability	X	Y
$(1-p)^2$	0	0
$p(1-p)$	1	0
p	2	1

The pair (X, Y) is associated since it has the same joint distribution as $(Y + Y \vee W, Y)$, where Y, W are i.i.d.

Now $P(X \neq Y) = p(2-p)$, while if U, V are independent, with U equal in distribution to X and V equal in distribution to Y, then $P(U \neq V) = p(3 - 4p + 2p^2)$. Thus, when $\frac{1}{2} < p < 1$, the independent variables have a smaller probability of being distinct than the dependent ones, in contrast to the result of Theorem 3.1. Thus, under any such choice of p, the pair (X, Y) is associated, but not affiliated.

In the following example, X and Y are identically distributed, and the pair is affiliated (and associated) if and only if $0 \leq p \leq \frac{4}{9}$.

probability	X	Y
p	0	0
$p/2$	0	1
$p/2$	1	0
$1-2p$	1	1

A sufficient condition for a pair X, Y to be affiliated is that it should be associated under every total ordering of S. To see this, let A be a subset of S. Given any element a in A we may choose a total order of S so that a is minimal and all elements of A^c are larger than all elements of A. The function

$$F(x,y) = 1_{A^c}(x)1_{\{a\}}(y)$$

is then non-decreasing in x and non-increasing in y, hence by association we have

$$P(X \notin A, Y = a) = EF(X,Y) \leq P(X \notin A)P(Y = a).$$

This implies that

$$P(X \notin A|\sigma(Y)) \leq P(X \notin A) \text{ on } \{Y \in A\},$$

hence the pair X, Y is affiliated.

Boolean associated pairs are automatically affiliated. It is shown in [1] that a Boolean pair X, Y is associated if and only if $1 - X, 1 - Y$ is associated. Since a 2 element set has only 2 possible total orders, the desired result then follows from the previous paragraph.

A method to produce a new affiliated pair U, V with a larger set of possible values than X, Y is to split certain atoms: consider an atom $\{X = a, Y = b\}$ and replace it with a pair of atoms $\{U = c, V = c\}$ and $\{U = d, V = d\}$, where c and d are new values not already taken on by X or Y. The original probability can be split between these atoms arbitrarily. Let $(U, V) = (X, Y)$ on the rest of the sample space, and then continue the same procedure until the value $Y = b$ has been eliminated. Let A be some set of values that may contain values from the original S as well as added values. Clearly $P(U \notin A, V = c) = 0$ whenever A contains c and c is one of the new values. If $e \neq b$ is a value of Y belonging to A, then $X = U$ on

$\{V = e\} = \{Y = e\}$. Let B be obtained from A by adjoining all values of X that were eliminated during the procedure. Then

$$P(U \notin A, V = e) = P(X \notin B, Y = e) \leq P(X \notin B)P(Y = e) \leq P(U \notin A)P(Y = e).$$

Starting with an independent pair, this method may be used to produce a variety of examples.

5 Discussion

We cannot pretend to be very happy with the results of section 3.1. Condition (3.1) seems to be much too strong, and is not clear there are any interesting examples that satisfy it other than those that can be obtained from constant sequences and independent sequences using the constructions indicated above.

A number of open questions remain. We have been unable to find an *identically distributed* associated collection that does not satisfy the birthday inequality of Theorem 3.2. Does any such collection exist? (The first example in section 4 shows that we can have a non-identically distributed associated pair for which the inequality of Theorem 3.1 fails.) We also showed in section 4 that an associated Boolean pair is necessarily affiliated. Does this extend to larger collections?

References

[1] J.D. Esary, F. Proschan, and D. W. Walkup, Association of Random Variables, with Applications, *Ann. Math. Statist.* **38**(1967),1466-1474.

[2] T.R. McConnell, An Inequality Related to the Birthday Problem, Preprint (2001)