# The Birthday Problem when Birthdays are Dependent (Followup)

Terry R. McConnell
Syracuse University

November 2, 2022

**Abstract**

This is a followup to our earlier paper, which showed 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

Let $X_1, X_2, \ldots, X_N$ be identically distributed random variables defined on a common probability space, and taking values in some finite set $S$. Let $Y_1, Y_2, \ldots, Y_N$ be independent variables with the same marginal distributions as the $X$ variables. In what follows, we shall say that the *birthday inequality* holds when

$$P(X_1, X_2, \ldots, X_N \text{ are distinct}) \leq P(Y_1, Y_2, \ldots, Y_N \text{ are distinct})$$

The example motivating the terminology arises when $S$ represents the days of the calendar year and the $X_i$ represent birthdays of people in a group of size $N$. The famous Birthday "Paradox" is based on the fact that the quantity on the right side is unexpectedly small for seemingly modestly large values of $N$, and our earlier work considered the extent to which the paradox perists, or is enhanced, when birthdays are suitably positively correlated. In particular, we considered association, a strong notion of positive correlation introduced in [1]. To our surprise, it is possible for associated sets of birthdays to be less likely to have some matching pair than independent ones. We then introduced an even stronger notion of positive correlation that is sufficient for the birthday inequality to hold.

## 2 Affiliated Random Variables

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

$$P(X_i \in A | \mathcal{F}_{\hat{i}}) \geq P(X_i \in A) \text{ on the event } \bigcap_{j \neq i} \{X_j \in A\}, \tag{2.1}$$

and also for subcollections.

Stated informally, the variables "want" to belong to sets the other variables belong to.

## 3 Examples

The following example shows that the birthday inequality may fail when variables are identically distributed and associated. ([2] had an example with variables that are not identically distributed.) It also provides an interesting example of an affiliated pair.

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

| X/Y | -1 | 0 | 1 |
|---|---|---|---|
| -1 | $\frac{p}{4}$ | $\frac{1-p}{4}$ | 0 |
| 0 | $\frac{1-p}{4}$ | $\frac{p}{2}$ | $\frac{1-p}{4}$ |
| 1 | 0 | $\frac{1-p}{4}$ | $\frac{p}{4}$ |

This distribution may be obtained as that of the position of a random walk after 2 steps when the $x$ and $y$ components perform simple symmetric random walk with step size $\frac{1}{2}$. The components are coupled by making the first steps equal, and the second steps equal with probability $p$.

The pair $(X, Y)$ is associated when $\frac{1}{3} \leq p$, affiliated when $\frac{1}{2} \leq p$, and the birthday inequality fails when $p < \frac{3}{8}$. For the last statement, let $U$ and $V$ be independent variables with the same distribution as $X$ (or $Y$.) Then $P(U = V) = \frac{3}{8}$. On the other hand, summing the diagonal entries of the table gives that $P(X = Y) = p$. Thus, when $p < \frac{3}{8}$, the independent variables have a larger chance of being equal than the dependent (positively correlated) ones, and in the range $\frac{1}{3} \leq p < \frac{3}{8}$ one has an example where the birthday inequality is violated by identically distributed associated variables.

The labor in showing that $(X, Y)$ is associated (or not) is eased by using Theorem 4.3 of [1]: To show the pair is associated, it suffices to show that for every value of $t$, the function $g(s) = P(X > t|Y = s)$ is non-decreasing. It is sufficient, of course, to consider two cases: $t = -1$ and $t = 0$. Other cases are automatic or follow from these. For $t = -1$ one computes, e.g., that $g(0) = P(X \in \{0, 1\}|Y = 0) = \frac{1/4+p/4}{1/2} = 1/2+p/2$. Similarly, $g(-1) = 1 - p$, and $g(1) = 1$. Thus, $g$ is non-decreasing if and only if $\frac{1}{3} \leq p$. When $t = 0$, the only possible value of $X$ is 1, and we find that $g(-1) = 0, g(0) = \frac{1-p}{2}, g(1) = p$. This is also non-decreasing if and only if $\frac{1}{3} \leq p$, establishing the claim.

For the statement about affiliation, note that (2.1) is automatic for any set containing $\{-1, 0, 1\}$ or disjoint from it. By symmetry, then, there are only 4 cases that need be tested for the choice of $A$: $\{0\}, \{1\}, \{0, 1\}$, and $\{-1, 1\}$. We have $P(X = 0|Y = 0) = p, P(X = 1|Y = 1) = p$. Also, $P(X \in \{0, 1\}|Y = 0) = \frac{p}{2}+\frac{1}{2}, P(X \in \{0, 1\}|Y = 1) = 1$. Finally, $P(X \in \{-1, 1\}|Y = 1) = p$. The result in every case exceeds the unconditional probability if and only if $p \geq \frac{1}{2}$.

We turn to an example that shows one method for constructing a new affiliated set $Y_1, Y_2, \ldots, Y_n$ from a given one, $X_1, X_2, \ldots, X_n$: pick an ordered pair $(j, k)$ of distinct indices and define $Y_j = X_k$, and $Y_i = X_i$ for $i \neq j$. If $i \neq j, k$, then (2.1) holds for the $Y's$ since the computation can be done as if we had simply dropped $X_j$ from the original collection. On the other hand, if $i = j$ then

$$P(Y_i \in A|\mathcal{F}_{\hat{i}}) = 1_{\{X_k \in A\}} = 1$$

on the event $\bigcap_{j \neq i}\{Y_j \in A\}$. The same holds when $i = k$.

It might seem that we could make the selection of a pair of indices at random, as long as it is done independently of all the $X's$, but this is not true. To see a very simple example of what can go wrong, suppose we start with a triple $(X, Y, Z)$, where the pair $X, Y$ is i.i.d Bernoulli (taking the values 0 and 1 with equal probability) and $Z$ is the non-random constant equal to 2. This triple is affiliated. Now toss a fair coin and make $X = Z$ if it is heads, and $Y = Z$ if it is tails. The new triple is not affiliated because the subcollection $(U, V)$ consisting of the first two variables is not. Its joint distribution is:

| U/V | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 0 | 0 | $\frac{1}{4}$ |
| 1 | 0 | 0 | $\frac{1}{4}$ |
| 2 | $\frac{1}{4}$ | $\frac{1}{4}$ | 0 |

We have $P(U = 0|V = 0) = 0$, while $P(U = 0) = \frac{1}{4}$. This example also shows that mixtures of affiliated sequences are not necessarily affiliated.

We turn now to a family of examples which allow for any number of variables $X_1, \ldots, X_N$ taking values in any finite set $S$. There is no harm in assuming $S$ is just the first $n$ natural numbers. Denote by $\mathbf{x}$ an $N$-tuple of elements of $S$, i.e., an element of the Cartesian product $S^N$. We will assign a probability $p(\mathbf{x})$ that is independent of the order of the components of $\mathbf{x}$, that is, if $\mathbf{y}$ is obtained by permuting the components of $\mathbf{x}$, then $p(\mathbf{y}) = p(\mathbf{x})$.

To be more precise, given any partition $N = d_1 + d_2 + \cdots + d_k$ of $N$ with $k$ parts arranged in non-increasing order, we shall say that an $N$-tuple $\mathbf{x}$ has type $[d_1, d_2, \ldots, d_k]$ if it has sets of $d_1, d_2, \ldots, d_k$ equal components, with distinct common values for each set. Thus, for example, the 10-tuple $5, 3, 1, 5, 3, 6, 6, 2, 2, 3$ has type $[3, 2, 2, 2, 1]$. Order the set $T$ of all possible types lexicographically, so that, e.g., the listed example is larger than any containing only ones and twos, and smaller than any example containing a 4 or higher digit. Explicitly, we have $[c_1, c_2, \ldots, c_m] < [d_1, d_2, \ldots, d_n]$ if and only if, for some $j$ we have $c_1 = d_1, \ldots, c_{j-1} = d_{j-1}, c_j < d_j$. We denote the type of a given $N$-tuple $\mathbf{x}$ by $[\mathbf{x}]$.

We will need below the following simple observation about lexicographic order. Suppose $c \geq c'$ represent the counts of some pair of distinct digits in a sequence $\mathbf{x}$, and $d \geq d'$ the same counts in another sequence $\mathbf{y}$. If $[c, c'] \geq [d, d']$, and the counts of all other digits are equal, then $[\mathbf{x}] \geq [\mathbf{y}]$.

Let $f : T \to \mathbb{R}_+$ be non-zero and non-decreasing relative to the order of $T$, and finally define $p(\mathbf{x}) = f([\mathbf{x}])/c$, with normalizing constant $c$ chosen so that this defines a probability on $S^N$.

Let $X_i$ be the usual i-th coordinate function on $S^N$, and we shall show that $X_1, X_2, \ldots, X_N$ is affiliated. In addition, the $X_i$ are exchangeable, since the type of a sequence is unchanged by permuting its terms; and the marginal distribution of each $X_i$ is uniform on $S$, since the distribution is unchanged by permutations of elements of $S$.

In what follows, we indicate individual elements of $S$ by lower case letters, and sequences of elements with boldface. Sequences are joined by concatenating their letters. We denote by $y(\mathbf{x})$ the number of times $y$ occurs in $\mathbf{x}$.

**Lemma 3.1.** *If $\mathbf{x}$ is a sequence of elements of $S$ of length $k < N$ and $u(\mathbf{x}) = 0$, then*

$$\sum_{\mathbf{y}} f([u\mathbf{x}\mathbf{y}]) \leq \sum_{\mathbf{y}} f([v\mathbf{x}\mathbf{y}]),$$

*where the sum extends over all possible sequences of length $N - k - 1$.*

Before beginning the proof, it should be noted that inequality does not in general hold between corresponding terms of the sums. For example, with $N = 4$, if $\mathbf{x} = v$ and $\mathbf{y} = uu$ then $[uvuu] = [3, 1] > [2, 2] = [vvuu]$.

Turning to the proof, we may suppose $u \neq v$. Let $a = v(\mathbf{x}), b = u(\mathbf{y})$. If $b < a$, then the frequency of $u$ does not determine the order position of the term belonging to $\mathbf{y}$ on either side, and then we clearly have the term-wise inequality $f([u\mathbf{x}\mathbf{y}]) \leq f([v\mathbf{x}\mathbf{y}])$. This same inequality must also hold if $a = b$: $u\mathbf{x}\mathbf{y}$ contains exactly $1 + a$ u's and $a + v(\mathbf{y})$ v's, while $v\mathbf{x}\mathbf{y}$ has $a$ u's and $1 + a + v(\mathbf{y})$ v's. The counts of all other elements are the same, so the inequality follows from the property of lexicographic order noted above. Thus we may restrict the summation to the set $B$ of $\mathbf{y}$ such that $b > a$. Assuming such a sequence $\mathbf{y}$, we can determine a sequence $\mathbf{z}$ up to permutation by the equation $[\mathbf{y}] = [u^a\mathbf{z}]$, where $u^a$ denotes $u$ repeated $a$ times, together with the requirement that $\mathbf{y}$ and $u^a\mathbf{z}$ be permutations of each other. Let $\mathbf{z}^*$ be the sequence obtained from $\mathbf{z}$ by changing every $u$ to a $v$ and *vice versa.*

To complete the proof, it suffices to show that

$$f([u\mathbf{x}u^a\mathbf{z}]) + f([u\mathbf{x}u^a\mathbf{z}^*]) \leq f([v\mathbf{x}u^a\mathbf{z}]) + f([v\mathbf{x}u^a\mathbf{z}^*]),$$

since the sum over $B$ may be written as the sum of such terms with $\mathbf{z} \neq \mathbf{z}^*$, plus half the sum of such terms with $\mathbf{z} = \mathbf{z}^*$.

To see this, let $c = u(\mathbf{z})$ and $d = v(\mathbf{z})$. The following ordered pairs then give for each of the 4 terms, $(\alpha, \beta)$, where $\alpha$ is the frequency of $u$ and $\beta$ is the frequency of $v$ in the sequence occurring in the given terms (all elements other than $u$ and $v$ occur with equal frequencies in all 4 sequences):

$$(a + c + 1, a + d), (a + d + 1, a + c), (a + c, a + d + 1), (a + d, a + c + 1).$$

If $c > d$, then inspection of these ordered pairs shows that no term can exceed the last term, and the third term is equal to the second; while if $d > c$ then no term can exceed the third, and the fourth is equal to the first. If $c = d$, then all 4 terms are equal.

To show that $X_1, \ldots, X_N$ is affiliated, i.e. that (2.1) holds for a given sub-collection of size $k$, it suffices, since the variables are exchangeable, to show that

$$P(X_1 \in A | X_2 \ldots X_k = \mathbf{x}) \geq P(X_1 \in A) \tag{3.1}$$

holds for any $\mathbf{x}$ having no component outside of $A$. Replacing $A$ temporarily by an arbitrary element $v$ of $A$, and multiplying by $P(X_2 \ldots X_k = \mathbf{x})$, the left side of (3.1) becomes

$$\sum_{\mathbf{y}} p(v\mathbf{xy}) = \frac{1}{c} \sum_{\mathbf{y}} f([v\mathbf{xy}]).$$

Let $u$ be any element not belonging to $A$. Then by Lemma 3.1 we have,

$$\sum_{\mathbf{y}} p(u\mathbf{xy}) \leq \frac{1}{c} \sum_{\mathbf{y}} f([v\mathbf{xy}]).$$

Summing over $u$ and then $v$ we obtain,

$$|A| P(X_1 \notin A | X_2 \ldots X_k = \mathbf{x}) P(X_2 \ldots X_k = \mathbf{x}) \leq (n - |A|) \sum_{\mathbf{y}} P(X_1 \in A | X_2 \ldots X_k = \mathbf{x}) P(X_2 \ldots X_k = \mathbf{x}),$$

where $|A|$ is the cardinality of $A$. From this, it follows by rearrangement that

$$P(X_1 \in A | X_2 \ldots X_k = \mathbf{x}) \geq \frac{|A|}{n} = P(X_1 \in A).$$

# 4 Discussion

We showed in [2] that if a pair of $S$ valued random variables is associated under every total ordering of $S$, then it is affiliated. The argument shows that for larger collections $X_1, X_2, \ldots, X_n$ associated under every total order of $S$, if $A$ is any subset of $S$, then $P(X_i \in A | X_j \in A, j \neq i) \geq P(X_i \in A)$, and also for sub-collections. This is weaker than affiliation since the conditioning event is not necessarily an atom of the sigma field generated by the variables involved. One also has a partial converse: If the collection is affiliated, then under any ordering of $S$ we have the probability inequality

$$P(X_1 > c_1, X_2 > c_2, \ldots, X_n > c_n) \geq P(X_1 > c_1) P(X_2 > c_2) \ldots P(X_n > c_n), \tag{4.1}$$

and also for sub-collections. By relabeling the variables if necessary, we may assume that the $c_i$ are non-increasing in $i$. Let $A = \{x \in S : x > c_n\}$. Then $\{X_1 > c_1, X_2 > c_2, \ldots X_{n-1} > c_{n-1}\} \subseteq \{X_j \in A, j \neq n\}$. Then

$$P(X_1 > c_1, X_2 > c_2, \ldots, X_n > c_n) = E(P(X_n > c_n | \mathcal{F}_{\hat{n}}), X_1 > c_1, \ldots X_{n-1} > c_{n-1}),$$

and, by affiliation,

$$E(P(X_n > c_n | \mathcal{F}_{\hat{n}}), X_1 > c_1, \ldots X_{n-1} > c_{n-1}) \geq P(X_1 > c_1) P(X_1 > c_1, X_2 > c_2, \ldots, X_{n-1} > c_{n-1}).$$

We may repeat the argument to establish (4.1), which is a weaker form of association, as noted in [1].

# 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, The Birthday Problem when Birthdays are Dependent, Preprint (2022)