8 Appendix

8.1 Details on the exact calculation of $p_{m,n,p,d}$

The plan of attack is as follows. Fix a p-subset of $\{X_1, \cdots, X_n\}$, and denote the ordered members as $x_1 \le x_2 \le \cdots \le x_p$. Let $\alpha_{m,p,d} = P(\text{span}\{x_1, \cdots, x_p\} \le d)$. If T denotes the number of B_i that occur, i.e., $T = \sum_{i=1}^{N} I_{B_i}$, then $E(T) = \lambda = N\alpha_{m,p,d}$, and we may, for instance, use the Poisson approximation $p_{m,n,p,d} = P(T > 0) \approx 1 - e^{-\lambda}$. A main task is to analyze $\alpha_{m,p,d}$.

To illustrate the counting, we take the case of our direct interest first, i.e., $m = 365, p = d = 4$. It will be necessary to also keep track of the spacings between the successive ordered values $x_1 \le x_2 \le \cdots \le x_p$. Thus, define $d_j = x_{j+1} - x_j$, and given k, k', δ where $1 \le k \le k', \delta \ge 1$, consider the sets

$$C_{k,k',\delta} = \{x_1 \le x_2 \le \cdots \le x_p : 1 \le x_i \le m\, \forall i,\ k \le d_j \le k'\, \forall j,\ \text{span}\{x_1, x_2, \cdots, x_p\} \le \delta\}.$$

$$C^0_{k,k',\delta} = \{x_1 \le x_2 \le \cdots \le x_p : 1 \le x_i \le m\, \forall i,\ k \le d_j \le k'\, \forall j,\ x_p - x_1 \le \delta\}.$$

$$C^1_{k,k',\delta} = \{x_1 \le x_2 \le \cdots \le x_p : 1 \le x_i \le m\, \forall i,\ k \le d_j \le k'\, \forall j,\ x_p - x_1 > \delta, \text{span}\{x_1, x_2, \cdots, x_p\} \le \delta\}.$$

Note that C^0, C^1 form a partition of C, i.e.,

$$C_{k,k',\delta} = C^0_{k,k',\delta} \cup C^1_{k,k',\delta}.$$

The second set in the partition arises because span defined in terms of the circular distance d can be small even when the Euclidean distance is large.

Note that since k in the above sets C, C^0, C^1 is taken to be at least one, for the members of $C_{k,k',\delta}$, the x_i are necessarily distinct. Since some of the x_i may actually coincide (i.e., people may have identical birthdays), we must also keep track of the number of distinct x_i and if some coincide, which ones do. Thus, define, with $p = 4$, also the events $D_i, i = 1, \ldots, 7$ representing all different possible cases of coincidence, and

$$E = \{\text{span}\{x_1, \cdots, x_4\} \le 4\},$$

$$E_0 = \{x_4 - x_1 \le 4\},\ \ E_1 = \{x_4 - x_1 > 4, \text{span}\{x_1, \cdots, x_4\} \le 4\}.$$

Once again, E_0, E_1 form a partition of $E : E_o \cup E_1 = E$. We now have,

$$\begin{aligned} \alpha_{m,p,d} &= P(C_{1,4,4}) + \textstyle\sum_{i=1}^{7} P(D_i \cap E) \\ &= P(C^0_{1,4,4}) + P(C^1_{1,4,4}) + \textstyle\sum_{i=1}^{7} P(D_i \cap E_0) + \sum_{i=1}^{7} P(D_i \cap E_1). \end{aligned} \tag{32}$$

For this example with $m = 365$ and $p = d = 4$, we have $\alpha_{m,p,d} = 7.584 \times 10^{-6}$. This value for $\alpha_{m,p,d}$ has been verified by a simulation of 5,000 simulation trials.

Details of all possible cases for coincidence, $D_i, i = 1, \ldots, 7$.

$$D_1 = \{x_1 = x_2 < x_3 < x_4\}$$

$$D_2 = \{x_1 = x_2 < x_3 = x_4\}$$

$$D_3 = \{x_1 < x_2 = x_3 < x_4\}$$

$$D_4 = \{x_1 < x_2 < x_3 = x_4\}$$

$$D_5 = \{x_1 = x_2 = x_3 < x_4\}$$

$$D_6 = \{x_1 < x_2 = x_3 = x_4\}$$

$$D_7 = \{x_1 = x_2 = x_3 = x_4\}$$

□

Details of calculating Equation (32). By Lagrange (1963), pp 41, formula (35), the number of p-combinations $x_1 \leq x_2 \leq \cdots \leq x_p$ in the set $C^0_{k,k',\delta}$ for general $1 \leq k \leq k', \delta$ equals

$$\begin{aligned} f_{m,p,k,k',\delta} = &\sum_{i \geq 0} (-1)^i \binom{p-1}{i} \frac{m - (p-1)(k-m+\delta) - i(k'-k+1)}{p} \\ &\times \binom{\delta - (k-1)(p-1) - i(k'-k+1)}{p-1}, \end{aligned} \tag{33}$$

where the range of summation extends over i such that $0 \leq i \leq p-1, \delta-(k-1)(p-1)-i(k'-k+1) \geq p-1$. For the special values $k = 1, \delta = k'$, this simplifies to the expression

$$f_{m,p,k'} = \frac{m + (p-1)(m-\delta-1)}{p} \binom{\delta}{p-1}. \tag{34}$$

With $m = 365, p = \delta = k' = 4, k = 1$, this equals 1445. The same formula applies to the number of points in $D_3 \cap E_0$ and $D_4 \cap E_0$, with p reduced to 3; the number of elements is 2170 in each of these two cases. By direct counting, the number of elements in $C^1_{1,4,4}$ is 14.

We have to now count the number of points in the remaining events $D_i \cap E_j, i = 1, 2, \cdots, 6, j = 0, 1$. For example, for $D_1 \cap E_0$, if x_1 is between 4 and 361, and $x_2 = x_1$, then the two distinct numbers $x_3 < x_4$ can be placed within a straight line distance of four in six ways; if $x_1 = x_2 = 362$ or 363, the number of ways to place x_3, x_4 reduces to 3 and 1 respectively. For $x_1 = 1, 2, 3$, the circular span can be ≤ 4 even if x_3, x_4 are sufficiently close to $m = 365$. Thus, these cases require boundary correction. By direct counting, the number of points in $D_1 \cap E_1$ when $x_1 \leq 3$ is 28, and thus the total number of points in $D_1 \cap E$ is $358 \times 6 + 3 + 1 + 28 = 2180$.

We provide a summary table of the number of points in the sets $D_i \cap E_j$:

	E_0	E_1	E
D_1	2148	32	2180
D_2	1434	26	1460
D_3	2170	22	2192
D_4	2170	22	2192
D_5	1434	26	1460
D_6	1434	26	1460
D_7	365	0	365

Thus, pooling all the counts, and accounting for multinomial reallocation, with $m = 365, p = d = 4$,

$$\alpha_{m,p,d} = \frac{(1445 + 14) \times 24 + 2180 \times 12 + 1460 \times 6 + 2 \times 2192 \times 12 + 2 \times 1460 \times 4 + 365}{365^4}$$

$$= 7.584 \times 10^{-6}. \tag{35}$$

□

Remark: It is interesting that a bound provided in Hunter (1976) also almost reproduces the value of $\alpha_{m,p,d} = 7.584 \times 10^{-6}$. The Hunter bound says that for k given events, $A_1, A_2, \ldots, A_k$,

$$P(\cup_1^k A_i) \leq \sum_1^k P(A_i) - \sum_2^k P(A_{(i)}A_i), \tag{36}$$

where (i) denotes some arbitrary choice of subscripts in $\{1, \ldots, i-1\}$ for $i > 1$. Let A_i denote the event that four birthdays all fall within the 5 days, $(i, i+1, i+2, i+3, i+4)$. The right hand side of (36) equals $365(5/365)^4 - 365(4/365)^4 = 7.588 \times 10^{-6}$.

8.2 Proofs of theorems in Section 3

Proof of Theorem 3.1. We include the proof only for completeness. We have,

$$P(x_1 \geq x, x_p \leq y) = \frac{(y - x + 1)^p}{m^p}, 1 \leq x \leq y \leq m.$$

This gives:

$$P(x_1 = x, x_p = y) = \frac{(y - x + 1)^p - 2(y - x)^p + (y - x - 1)^p}{m^p}, 1 \leq x \leq m-1, x+1 \leq y \leq m,$$

and

$$P(x_1 = x_p = x) = \frac{1}{m^p}, 1 \leq x \leq m.$$

Hence,

$$P(x_p - x_1 \le d) = \sum_{j=0}^{d} P(x_p - x_1 = j) = \sum_{j=1}^{d} \sum_{x=1}^{m-j} P(x_1 = x, x_p = x + j) + \frac{m}{m^p}$$

$$= \frac{m + \sum_{j=1}^{d} (m-j) \Big[(j+1)^p - 2j^p + (j-1)^p \Big]}{m^p}$$

$$= \frac{m + m \sum_{j=1}^{d} \Big[(j+1)^p - 2j^p + (j-1)^p \Big] - \sum_{j=1}^{d} \Big[j(j+1)^p - 2j^{p+1} + j(j-1)^p \Big]}{m^p}$$

$$= \frac{m \Big[\sum_{j=1}^{d+1} j^p + \sum_{j=1}^{d-1} j^p - 2 \sum_{j=1}^{d} j^p \Big] + 2 \sum_{j=1}^{d} j^{p+1} - \sum_{j=1}^{d+1} j^{p+1} - \sum_{j=1}^{d-1} j^{p+1} + d^p + (d+1)^p}{m^p}$$

$$= \frac{(m+1)(d+1)^p - (m-1)d^p + d^{p+1} - (d+1)^{p+1}}{m^p}$$

$$= \frac{(m-1-d) \Big[(d+1)^p - d^p \Big] + (d+1)^p}{m^p}, \tag{37}$$

as stated in the theorem. □

8.3 Poisson and Some Other Approximations

A large simulation conducted by us gave the Monte Carlo estimate of .099 for the probability that in a group of 30, at least four with birthdays within four days of each other exist. From the value of $\alpha_{m,p,d}$, the expected number T of quadruplets in a group of 30 with birthdays within at most four days of each other is

$$\lambda = N\alpha_{m,p,d} = \binom{30}{4} \times 7.584 \times 10^{-6} = .20784$$

An immediate theoretical upper bound on $P(T > 0)$ is

$$P(T > 0) \le E(T) = .20784.$$

A formal Poisson approximation gives

$$P(T > 0) \approx 1 - e^{-\lambda} = .18767,$$

of which,

$$P(T = 1) \approx \lambda\, e^{-\lambda} = .16884;\ P(T = 2) \approx \frac{\lambda^2\, e^{-\lambda}}{2} = .0175;\ P(T = 3) \approx \frac{\lambda^3\, e^{-\lambda}}{6} = .0012.$$

We can see that the Poisson approximation is not accurate.

If we adopt the heuristic value for the variance $\text{Var}(T) = E(T) = \lambda$, then we can obtain various lower bounds on $P(T > 0)$. For example, by the Alon-Spencer inequality (Alon and Spencer, 2000),

$$P(T = 0) \leq \frac{\text{Var}(T)}{E(T^2)} = \frac{\lambda}{\lambda + \lambda^2} = .82792,$$

and hence, $P(T > 0) \geq 1 - .82792 = .17208$, which comes close to the Poisson approximation value .18767.

We can also use the representation $T = \sum_{i=1}^{N} I_{B_i}$, to get

$$E(T^2) = E\Big[\sum_{i=1}^{N} I_{B_i} + 2\sum\sum_{j>i} I_{B_i} I_{B_j}\Big] = \lambda + 2\sum\sum_{j>i} P(B_i \cap B_j),$$

and with still the heuristic value $\text{Var}(T) = E(T) = \lambda$,this results in $2\sum\sum_{j>i} P(B_i \cap B_j) = \lambda^2$. Therefore, using the Móri-Székely inequality (Móri and Székely (1985)),

$$P(T = 1) \geq P(\cup_{i=1}^{N} B_i) - 2\sum\sum_{j>i} P(B_i \cap B_j) = \lambda - \lambda^2 = .16464,$$

and this comes close to the Poisson approximation $P(T = 1) \approx .16884$. However, each of these approximations overestimates the true value in this specific example.

8.4 The Case of a Random Group Size

In our illustrative example, the number of mass shootings is of course not a constant from year to year. In the US, it seems to average out to about one in two weeks. This would suggest a Poisson process with a rate of 25 to 30, with a year as unit time. The same is true of other processes one may naturally monitor, such as accidents, hurricanes and tornadoes, terrorist attacks, etc.

Since $\alpha_{m,p,d}$ does not depend on the group size, we have, if the group size $n \sim F$,

$$E[T] = E_F\left[\binom{n}{p}\right] \alpha_{m,p,d}.$$

If $F = \text{Poisson}(c)$, then it is well known that $E_F\left[\binom{n}{p}\right] = \frac{c^p}{p!}$, and hence, $E[T] = \frac{c^p}{p!}\alpha_{m,p,d}$. In particular, if $m = 365, p = d = 4$ and $c = 30$, then we get $E[T] = .25596$, which is greater than $\lambda = .20784$, the expected value of T when n is held constant at 30. The increase occurs because of the right tail of the Poisson group size. A Poisson approximation to the distribution of T is less credible when the group size is

random. However, even with a random group size, we can still apply the theoretical upper bound

$$P(T > 0) \leq E(T) = .25596.$$