An Asymptotic Approximation for the Birthday Problem

S. Ejaz Ahmed and Richard J. McIntosh

ABSTRACT. It is known that for a class of 23 students the probability that at least two students have the same birthday is more than 0.5. Suppose that the number of days in the calendar tends to infinity. For a fixed number p with $0 < p < 1$ we give an asymptotic formula and a simple proof, not using Stirling's formula, for the minimum class size to ensure a probability of at least p that two or more students have the same birthday.

There were two U.S. Presidents born on November 2 — James K. Polk, 1795, and Warren G. Harding, 1865. Three U.S. Presidents died on the 4^{th} of July — John Adams, 1826, Thomas Jefferson, 1826, and James Munroe, 1831. Given a small collection of people there is good chance that two or more individuals will have the same birthday. It is not difficult to show that 23 is the minimum number of students required in a class to ensure a probability of at least 0.5 that two or more students will have the same birthday (see for example, section II.3, p. 33 of W. Feller, An Introduction to Probability Theory and Its Applications, vol. 1, Wiley, New York, 1968).

The number of ways of choosing a sequence of k days from a calendar with n days is n^k because for each day selected we have n choices. If we require our choice of k days to be distinct, then the number of ways of doing this is reduced to $n(n-1)(n-2)\cdots(n-(k-1))$ because in our selection process we must avoid the days already chosen. Therefore if k days are chosen at random, then the probability that they will be distinct is equal to

$$\frac{n(n-1)(n-2)\cdots(n-(k-1))}{n^k}.$$

We will let the k days chosen be the birthdays of students in a class of size k. So the probability of two or more students having the same birthday is equal to

$$1 - \frac{n(n-1)(n-2)\cdots(n-(k-1))}{n^k}. \tag{1}$$

Now let $0 < p < 1$ and define k (as a function of n) to be the minimum class size to ensure a probability of at least p that two or more students will have the same birthday. We immediately see that k is the smallest positive

integer making (1) greater than or equal to p. It follows that for this definition of k,

$$\lim_{n\to\infty}\left(1-\frac{n(n-1)(n-2)\cdots(n-(k-1))}{n^k}\right) = p\,,$$

or equivalently,

$$\lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-(k-1))}{n^k} = 1-p\,. \tag{2}$$

In working with limits as $n\to\infty$ the concept of asymptotic functions is very useful. Two functions $f(n)$ and $g(n)$ are said to be *asymptotic* if

$$\lim_{n\to\infty}\frac{f(n)}{g(n)} = 1\,,$$

in which case we write $f(n)\sim g(n)$. With this notation in mind we now state our main theorem:

Theorem. *Fix $0<p<1$ and let the number of days n in the calendar tend to infinity. Then the minimum class size to ensure a probability of at least p that two or more students will have the same birthday is given asymptotically by*

$$k\sim\sqrt{2n\ln\frac{1}{1-p}}\,. \tag{3}$$

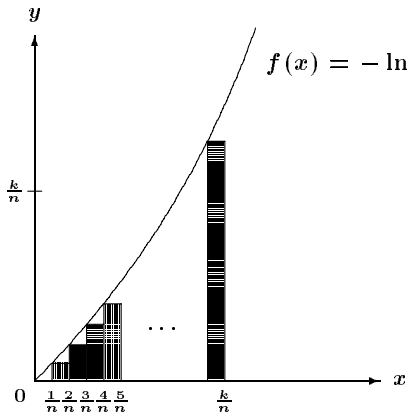
Proof. It is clear that as n tends to infinity so does k. Equation (2) can be rewritten in the form

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)=1-p\,.$$

Taking logarithms and multiplying by $-\frac{1}{n}$ we obtain

$$\sum_{j=0}^{k-1}\frac{1}{n}\left(-\ln\left(1-\frac{j}{n}\right)\right) \sim -\frac{1}{n}\ln(1-p)\,. \tag{4}$$

The sum on the left side of (4) is a lower Riemann sum for the function $f(x)=-\ln(1-x)$ on the interval $[0,\,\frac{k}{n}]$.



From the right side of (4) we see that the value of this Riemann sum is asymptotic to $-\frac{1}{n}\ln(1-p)$, which tends to 0 as $n \to \infty$. This implies that $k/n \to 0$ as $n \to \infty$. Since the slope $f'(0) = 1$, we see that as $n \to \infty$ the graph of the Riemann sum approximates the shape of a 45 degree right triangle with legs $\frac{k}{n}$ and area $-\frac{1}{n}\ln(1-p)$. By the well-known formula for the area of a triangle, it follows that

$$\frac{1}{2}\left(\frac{k}{n}\right)^2 \sim -\frac{1}{n}\ln(1-p)\,.$$

Therefore

$$\frac{k}{n} \sim \sqrt{-\frac{2}{n}\ln(1-p)}$$

and so

$$k \sim \sqrt{2n\ln\frac{1}{1-p}}\,,$$

which completes the proof.

The difference between k and our asymptotic approximation of k is not uniformly bounded for $0 < p < 1$ because the logarithm term in (3) tends to infinity as $p \to 1$, but $k \leq n+1$. To explore the behaviour of this asymptotic approximation we have computed its value for certain values of p and n, listed in the tables below. In the situations illustrated in these tables we see that

$$\left| \left\lceil \sqrt{2n\ln\frac{1}{1-p}} \right\rceil - k \right| \leq 1\,, \tag{5}$$

where $\lceil x \rceil$ denotes the smallest integer $\geq x$. Further calculations suggest that (5) holds for $0 < p \leq 0.98$ and $n \geq 1$.

Acknowledgement. Support by the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

Table I: $p = 0.1$

n	k	$\sqrt{2n \ln \frac{1}{1-p}}$
1	2	0.459
2	2	0.649
3	2	0.795
4	2	0.918
5	2	1.026
6	2	1.124
7	2	1.215
8	2	1.298
9	2	1.377
10	2	1.452
20	3	2.053
30	4	2.514
40	4	2.903
50	4	3.246
60	5	3.556
70	5	3.841
80	5	4.106
90	5	4.355
100	6	4.590
200	7	6.492
300	9	7.951
365	10	8.770
400	10	9.181
500	11	10.265
600	12	11.244
700	13	12.145
800	14	12.984
900	15	13.771
1000	15	14.516
2000	21	20.529
5000	33	32.459
10000	47	45.904
20000	66	64.919
50000	104	102.645
100000	146	145.162
200000	206	205.291
500000	326	324.593
1000000	460	459.044

Table II: $p = 0.5$

n	k	$\sqrt{2n \ln \frac{1}{1-p}}$
1	2	1.177
2	2	1.665
3	3	2.039
4	3	2.355
5	3	2.633
6	4	2.884
7	4	3.115
8	4	3.330
9	4	3.532
10	5	3.723
20	6	5.266
30	7	6.449
40	8	7.447
50	9	8.326
60	10	9.120
70	11	9.851
80	11	10.531
90	12	11.170
100	13	11.774
200	17	16.651
300	21	20.393
365	23	22.494
400	24	23.548
500	27	26.328
600	30	28.841
700	32	31.151
800	34	33.302
900	36	35.322
1000	38	37.233
2000	53	52.655
5000	84	83.255
10000	119	117.741
20000	167	166.511
50000	264	263.277
00000	373	372.330
200000	527	526.554
500000	833	832.555
1000000	1178	1177.410

Table III: $p = 0.9$		
n	k	$\sqrt{2n \ln \frac{1}{1-p}}$
1	2	2.146
2	3	3.035
3	4	3.717
4	4	4.292
5	5	4.799
6	5	5.257
7	6	5.678
8	6	6.070
9	7	6.438
10	7	6.786
20	10	9.597
30	12	11.754
40	14	13.572
50	15	15.174
60	17	16.623
70	18	17.954
80	19	19.194
90	21	20.358
100	22	21.460
200	31	30.349
300	37	37.169
365	41	40.999
400	43	42.919
500	48	47.985
600	53	52.565
700	57	56.777
800	61	60.697
900	65	64.379
1000	68	67.861
2000	96	95.971
5000	152	151.743
10000	215	214.597
20000	304	303.485
50000	480	479.853
100000	679	678.614
200000	960	959.705
500000	1518	1517.427
1000000	2146	2145.966

S. Ejaz Ahmed
Department of Mathematics
and Statistics
University of Regina
Regina, Saskatchewan
Canada S4S 0A2
ahmed@math.uregina.ca

Richard J. McIntosh
Department of Mathematics
and Statistics
University of Regina
Regina, Saskatchewan
Canada S4S 0A2
mcintosh@math.uregina.ca