DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN

Madison, Wisconsin 53706

TECHNICAL REPORT NO. 591

November, 1979

THE BIRTHDAY PROBLEM WITH UNEQUAL PROBABILITIES

by

Jerome Klotz[1]
University of Wisconsin, Madison

[1] This research was supported in part by NIH Grant 5 R01 CA 18332-04 and U.S. Army Contract DAAG29-75-C-0024.

AMS 1979 subject classification. Primary 60C05; Secondary 6001, 6004

Key Words and Phrases: Birthday problem, unequal cell probabilities, symmetric functions, partitions, matching probability

THE BIRTHDAY PROBLEM WITH UNEQUAL PROBABILITIES[1]

by

Jerome Klotz
University of Wisconsin at Madison

The probability that r people have all birthdays different is estimated for $r = 1(1)25$. A representation in terms of symmetric functions makes the computation feasible and birthdays from 41,208 people are used to estimate the unequal probabilities of daily birth. Using these estimates the probabilities of no repetition are computed.

1. Introduction. Introductory probability books often consider the problem of matching and illustrate it by calculating the probability of no repeated birthdays in various sized groups (e.g. Feller (1957) p. 31, 32 or Parzen (1960) p. 46, 47). For N categories and r items selected at random with replacement (e.g. r birthdays, N = 365)

$$P(A_r) = (N)_r/N^r = N(N-1)\dots(N-r+1)/N^r \tag{1.1}$$

is the probability of the event A_r of no repeated categories among r items assuming equiprobable selection (probability 1/N). For N = 365, $P(A_r) > 1/2(<1/2)$ for $r < 23(\geq 23)$ and bettors have often taken advantage of this unintuitive result.

In reality there are 366 possible birthdays (including February 29) and natality differs with day of the year as well as geographic location. The question arises how this affects the odds.

2. The unequal probability case. Let p_k, $k = 1,2,\dots,N$ be the probability of selection from category k. Then for this case

$$P(A_r) = r! \sum_{1 \le i_1 < i_2 \ldots < i_r \le N} \sum \cdots \sum p_{i_1} p_{i_2} \ldots p_{i_r} \tag{2.1}$$

where the sum is over all possible subscripts $i_1, i_2, \ldots, i_r$ taking integer values with $1 \le i_1 < i_2 < \ldots < i_r \le N$. The total of $\binom{N}{r}$ such terms can be exceedingly large. For $N = 366$, $r = 23$ the total is exactly

17 41860 69523 59682 84110 50555 87492 48400

and practical calculation using (2.1) is impossible even on the fastest computers.

3. <u>A symmetric function representation.</u> The symmetric functions

$$a_1 = \sum_{i=1}^{N} p_i \qquad \text{and} \qquad S_1 = \sum_{i=1}^{N} p_i$$

$$a_2 = \sum_{1 \le i < j \le N} \sum p_i p_j \qquad\qquad S_2 = \sum_{i=1}^{N} p_i^2$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$a_r = \sum_{1 \le i_1 < i_2 < \ldots < i_r \le N} \sum \cdots \sum p_{i_1} p_{i_2} \ldots {}_{i_r} \qquad S_r = \sum_{i=1}^{N} p_i^r$$

can be related to each other (Girard (1629) and MacMahon (1915) p. 6). The expression in terms of $S_1, S_2, \ldots, S_r$ is

$$P(A_r) = r!a_r = r! \sum_{\substack{0 \le t_1, t_2, \ldots, t_r \le r \\ \sum_j jt_j = r}} \sum \cdots \sum (-1)^{r + \Sigma t_j} \left(\prod_{j=1}^{r} (S_j/j)^{t_j} / t_j! \right) \tag{3.1}$$

and can be derived from the generating function for a_r as follows:

$$1-a_1x + a_2x^2-\dots+(-1)^N a_N x^N = (1-p_1x)(1-p_2x)\dots(1-p_nx)$$

$$= \exp[\sum_{i=1}^{N} \ln(1-p_ix)] = \exp[-\sum_{i=1}^{N}\sum_{j=1}^{\infty}(p_ix)^j/j]$$

$$= \exp[-\sum_{j=1}^{\infty} S_jx^j/j] = \prod_{j=1}^{\infty}[\sum_{t_j=0}^{\infty}(-1)^{t_j}(S_jx^j/j)^{t_j}/t_j!]$$

$$= 1 + \sum_{r=1}^{N}[\sum_{\substack{0\le t_1,t_2,\dots,t_r\le r \\ \sum_j jt_j = r}}\sum\dots\sum(-1)^{\Sigma t_j}\prod_{j=1}^{r}(S_j/j)^{t_j}/t_j!]x^r .$$

Multiplication of exponential series gives the last expression and equating coefficients of x^r gives (3.1). The computation of (3.1) is illustrated for $r = 6$ in Table 1 and involved the generation of all possible partitions of r.

A considerable computational reduction occurs using (3.1) vs. (2.1). In particular (3.1) has only 1255 terms for $r = 23$ as indicated by tables of the number of unordered partitions in Abramowitz and Stegun (1944) p. 836.

TABLE 1

Systematic Computation of $P(A_r)$ for $r = 6$.

Partitions of r $V = (V_1, V_2, \ldots, V_r)$	$\tilde{t} = (t_1, t_2, \ldots, t_r)$ where $t_j = \#(V_i = j)$	$r!(-1)^{r+\Sigma t_j} / \prod_{j=1}^{r} (j)^{t_j} t_j!$
1 1 1 1 1 1	6 0 0 0 0 0	1
2 1 1 1 1 0	4 1 0 0 0 0	- 15
2 2 1 1 0 0	2 2 0 0 0 0	45
2 2 2 0 0 0	0 3 0 0 0 0	- 15
3 1 1 1 0 0	3 0 1 0 0 0	40
3 2 1 0 0 0	1 1 1 0 0 0	-120
3 3 0 0 0 0	0 0 2 0 0 0	40
4 1 1 0 0 0	2 0 0 1 0 0	- 90
4 2 0 0 0 0	0 1 0 1 0 0	90
5 1 0 0 0 0	1 0 0 0 1 0	144
6 0 0 0 0 0	0 0 0 0 0 1	-120

$$P(A_6)=S_1^{6}-15S_1^{4}S_2+45S_1^{2}S_2^{2}-15S_2^{3}+40S_1^{3}S_3-120S_1S_2S_3+40S_3^{2}-90S_1^{2}S_4+90S_2S_4+144S_1S_5-120S_6$$

4. Estimating unequal birth probabilities. The number of births on a given day of a year depends on many factors. Some of these factors are geographic location, economic conditions, weather, war, random events (e.g. power failures in New York, blizzards in Chicago), and medical practice. Rindfuss et al (1979) give convincing evidence that physician convenience is an important factor in dates of birth with Sundays and holidays having fewer deliveries.

A reasonably representative collection of birthdays to estimate p_i, $i = 1,2,\ldots,366$ was obtained from 41,208 Wisconsin residents who died in 1975. Table 2 gives numbers by day for each month.

TABLE 2

Birthdays of 41,208 Wisconsin Residents

Day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1	142	107	131	118	106	121	110	121	119	104	128	102	
2	95	155	125	112	104	104	107	99	108	106	110	107	
3	100	114	123	136	105	99	107	99	117	111	98	100	
4	126	119	128	126	139	106	138	115	116	104	116	99	
5	108	117	103	111	131	98	110	108	131	122	94	129	
6	114	116	134	122	120	111	110	109	123	120	114	112	
7	94	100	111	103	96	108	123	95	127	113	113	96	
8	119	116	138	116	115	112	96	137	117	108	105	111	
9	121	111	121	124	102	116	92	102	116	105	101	91	
10	108	108	113	110	115	107	116	112	99	132	118	109	
11	99	101	102	138	102	106	99	105	81	110	131	90	
12	129	130	128	114	97	123	125	131	132	94	115	117	
13	108	115	117	112	115	112	104	109	100	95	100	101	
14	117	148	115	132	115	113	108	107	121	123	115	103	
15	129	133	136	118	126	113	124	115	134	119	86	107	
16	140	101	103	124	119	108	130	137	119	125	93	109	
17	116	129	130	131	117	104	103	101	115	109	104	116	
18	113	101	123	124	102	116	86	115	119	99	102	130	
19	105	105	127	104	113	101	117	118	117	84	101	120	
20	106	139	130	105	121	104	120	128	123	119	98	121	
21	90	100	119	126	110	104	112	107	116	105	102	110	
22	113	139	130	138	106	113	125	126	97	99	111	102	
23	94	103	112	118	112	107	115	106	114	103	103	106	
24	119	117	138	101	106	117	109	114	132	122	100	143	
25	113	113	123	126	98	100	91	115	125	93	105	127	
26	73	113	121	112	122	99	128	112	122	109	104	74	
27	127	106	124	118	106	130	118	111	112	103	119	91	
28	117	147	112	105	132	111	115	116	127	107	103	93	
29	109	30	123	107	107	122	123	112	100	103	100	99	
30	121		109	111	121	119	108	106	111	89	99	83	
31	113		122		99		107	107		96		123	
Monthly Total	3478	3333	3771	3542	3479	3304	3476	3495	3490	3331	3188	3321	41,208

Estimates $p_i = x_i/\sum_j x_j$ where x_i is the number born on the i^{th} day were used to compute $S_1, S_2, \ldots, S_r$.

5. Probabilities for unequal cell probabilities. Using probability estimates calculated from Table 2, Table 3 gives $P(A_r)$ for $r = 1(1)25$. Included for comparison is the probability $(365)_r/365^r$ for the case $p_i = 1/365$ and the number of partitions of r. Calculations were carried out on a DEC PDP 11/70 at the University of Wisconsin and are believed accurate to a couple of units in the last decimal.

Thus there appears to be only 2 or 3 units difference in the 3rd decimal from the equiprobable case around the break even point of $r = 23$. It is conjectured that the combination of the additional day (February 29th) counteracts the decrease due to variation in the p_i Munford (1977), Bloom (1973). The main difficulty in calculating with (3.1) is the orderly generation of partitions. Lehmer, in Beckenbach (1964) p. 25, discusses one possible algorithm for their generation.

6. Acknowledgments. Interest in this problem originated at a dinner honoring retiring statistics chairman Gouri Bhattacharyya. A nickel was won by Barbara Klotz from George Box on her bet of no repeated birthdays despite a total of 30 in the group.

Thanks go to Steve Dahlberg as well as the Wisconsin Division of Health, Bureau of Health Statistics for providing the data in Table 2. Conversations with James Sweet and Ron Rindfuss were also helpful.

TABLE 3

Probability of no repeated birthdays in a group of size r.

r	$P(A_r)$	$(365)_r/365^r$	No. of partitions of r
1	1.0000	1.0000	1
2	.9972	.9973	2
3	.9917	.9918	3
4	.9835	.9836	5
5	.9726	.9728	7
6	.9591	.9595	11
7	.9432	.9438	15
8	.9249	.9257	22
9	.9044	.9054	30
10	.8819	.8831	42
11	.8575	.8588	56
12	.8314	.8331	77
13	.8038	.8056	101
14	.7749	.7769	135
15	.7449	.7471	176
16	.7140	.7164	231
17	.6824	.6850	297
18	.6503	.6531	385
19	.6180	.6209	490
20	.5856	.5886	627
21	.5533	.5563	792
22	.5212	.5243	1002
23	.4896	.4927	1255
24	.4586	.4617	1575
25	.4283	.4313	1958

REFERENCES

[1] Abramowitz, M. and Stegun, L.A. (1944). Handbook of Mathematical Functions, National Bureau of Standards Applied Math Series 55.

[2] Beckenbach, F. (1964). Applied Combinational Mathematics, John Wiley, New York.

[3] Bloom, D.M. (1973). A birthday problem, American Math. Monthly, 80, 1114-2.

[4] Feller, W. (1957). An Introduction to Probability Theory and Its Applications, Vol. 1, John Wiley, New York.

[5] Girard (1629). Invention Nouvelle en l'Algèbre. Amsterdam.

[6] MacMahon, P.A. (1915). Combinatory Analysis, Cambridge University Press, also Chelsea, New York, 1960.

[7] Munford, A.G. (1977). A note on the uniformity assumption in the birthday problem, American Statistician, 31, 119.

[8] Parzen, E. (1960). Modern Probability Theory, John Wiley, New York.

[9] Rindfuss, R.R., Ladinsky, J.L., Coppock, E., Marshall, V.W., and Macpherson, A.S. (1979). Convenience and the occurrence of births, Induction of Labor in the United States and Canada, International Journal of Health Services, 9, 439-460.