SHORT COMMUNICATION

ANALYSING THE APPROXIMATION MODEL TO BIRTHDAY PROBLEM

*CHOJI, D.N. & DEME, A.C.

Department of Mathematics
University of Jos, Nigeria
*(Corresponding author)
chojid@yahoo.com

Feller (1968) stated that the mathematical theory of probability gains practical value and an intuitive meaning in connection with real or conceptual experiments such as tossing a coin 100 times, throwing three dice, frequency of accidents, or determining from a group of people those with the same birthdays. All these descriptions are rather vague, and, in order to render the theory meaningful, we have to agree on what we mean by possible results of experiment or observation in question.

Consider an experiment of determining from a group of people, the probability of two people in the group having the same birthday. We will start by first assuming that there are 365 days in a year if leap years are neglected. This assumption is good enough as it takes four years before a leap year is realized and on condition that such a year is also not divisible by 100. Secondly, that each day has an equal chance of being a birthday. If we have for instance a group of say 370 people, it will be certain (a probability of one) that at least two people will have the same birthday since there are only 365 possible birthdays to go around. However, if on the other hand, there were only 2 people in the group, the chances that those 2 people share the same birthday to be quite small (a probability close to 0). But when the number in a group increases upward from 2 the probability increases that at least 2 people share a common birthday. The assumptions simplify the theory without affecting its applicability.

The history of probability (and of mathematics in general) showing a stimulating interplay of theory and applications; theoretical progress opens new fields of applications, and in turn applications lead to new problems and fruitful research.

The theory of probability is now applied in many diverse fields and flexibility of a general theory is required to provide appropriate tools for so great a variety of needs (Feller, 1968; Levine & Burke, 1972). One of the areas probability can be found useful is that of birthday problems. We shall be concerned with how combinations and permutations are put to use, and the error analysis involved. A simple program in the appendix has been developed to predict many situations with regards to birthdays.

MATHEMATICAL DERIVATION

Taking a random sample of size say *r* with replacement from a population of size *n* will result in n^r possible ways. In order to obtain the probability of the event that no element appears twice (equivalent to sampling without replacement we obtain *n* permutation *r*, that is ${}^n p_r$. If we assume that all arrangements have equal probability, we obtain the probability of no repetition in our sample is

$$P(X = r) = {}^n p_r = \frac{n(n-1)...(n-r+1)}{n^r} \quad \text{... (1)}$$

Such that for our birthday problem we assume that the years are of equal length that is n=365 days neglecting the February 29 of leap years. Secondly we assume that the birth rates are constant throughout the year. We can then obtain the probability that all r birthdays are different equals

$$P(x = r) = \frac{{}^{365}P_r}{365^r} = \frac{365}{365}.\frac{364}{365}.\frac{363}{365}...\frac{365-(r-1)}{365} \quad \text{... (2)}$$

which can be expressed as

$$P(x = r) = \frac{{}^{365}p_r}{365^r} = (1-\frac{1}{365})(1-\frac{2}{365})(1-\frac{3}{365})...(1-\frac{(r-1)}{365}) \quad \text{... (3)}$$

$$p = 1-\frac{{}^{365}p_r}{365^r} = 1-(1-\frac{1}{365})(1-\frac{2}{365})(1-\frac{3}{365})...(1-\frac{(r-1)}{365}) \quad \text{... (4)}$$

Equation 4 depicts the probability that at least two people share a birthday and so we use small p for the case of sharing the same birthday. We can derive a good numerical approximation to *p* when the *r* is small. This can be done by neglecting all cross products (what this means is that $1/365$ through $(r-1)/365$ are small such that the product of any two of them is very small, for example

$(1/365)*((r-1)/365)$, which is a small value.

In essence

$$\prod_{i=1}^{n}(1-a_i) \approx 1-\sum_{i=1}^{r-1}1*1*1*...a_i \quad \text{... (5)}$$

That is neglecting terms shown in Equations 6 to 8

$$\sum_{i=1}^{r-2}\sum_{i<j=2}^{r-1}1*1*1*...a_i a_j \quad \text{... (6)}$$

which is positive in value

$$\sum_{i=1}^{r-3}\sum_{i<j=2}^{r-2}\sum_{j<k=3}^{r-1}1*1*1*...a_i a_j a_k \quad \text{... (7)}$$

negative in value

$$\sum_{i=1}^{r-4}\sum_{i<j=2}^{r-3}\sum_{j<k=3}^{r-2}\sum_{k<l=4}^{r-1}1*1*1*...a_i a_j a_k a_l \quad \text{... (8)}$$

positive in value and the like. By this we will have that

$$\begin{aligned} P(x=r) &\approx 1-[1+2+3+...+(r-1)]/365 \\ &= 1-r(r-1)/(2(365)) \end{aligned} \quad \text{... (9)}$$

Since the sum of the first *r*-1 terms in a series in Arithmetic Progression is easily derivable to be *(r*-1*)r*/2. This implies that the probability that at least two people in a group of *r* people will have the same birthday will be p ≈ r(r – 1)/(2(365))

Suppose we would like to know the maximum *r* such that *p*<1/2 (case of median). Since there are 365 possible birthdays, it will be tempting to suggest that we would need just about half this number, which is 183. However, we require *r*=23 for the above to happen (Feller, 1968; Ross, 1976; Snell, 1987). To show this, we observed the probability *p* using Equation 2 that in a group of *r* people *r*=23 people, there is no duplication of birthdays, will be less than one half, that is

$$\frac{1}{2} \geq \frac{^{365}P_{23}}{365^{23}} = (1-\frac{1}{365}).(1-\frac{2}{365}).(1-\frac{3}{365})...(1-\frac{(r-1)}{365})$$

and taking logs of both sides and using the fact that log(1-*x*) ≈ -*x* for small *x*, we have

log(2) ≤ 1/365 +2/365 + . . . + (*r*-1)/365 = *r*(*r*-1)/730 which has a solution r=23.

Equations 3 and 9 give close results that for a given number people having different birthdays. For example when *r*=10 Equation 3 gives *p*=0.883 but Equation 9 gives *p*=0.877, a difference in value of 0.014.

However when *r* is becoming larger an approximation can be obtained by Equation 11.

$$\log(p) \approx [\,(1+2+3+...+(r\text{-}1)\,]/365 = r(r\text{-}1)/730 \quad \text{... (11)}$$

Such that for *r*=30 the Equation 11 gives the probability of 0.3037 where as by Equation 3 *p*=0.294. A better approximation when *r* is becoming bigger is we increase more terms of the expansion, e.g., pair cross products, triple cross products e.t.c.

ERROR ANALYSIS

Our concern will be to see how the approximation model will tend to differ from the theoretical probability model. We will study the situation that at least two people will share a given birthday, that is to say a given number people will not all have different birthdays. The diagram in Fig. 1 shows how Equations 4 (the birthday model) p_1 and Equation 10 (the approximation model) p_2 display the probabilities that for a given

number of people with at least two of them sharing a birthday and the error incurred d1 as a result of the approximation. The approximation in Equation 10 that is p_2 is good enough when dealing with groups of people below 16. Fig. 2 shows that as the number of people in a group becomes bigger, it will require that we increase more terms of the expansion in Equation 4 for a better approximation, that is it will require us to consider Equations 5, 6, 7, e.t.c. Table 1 displays the various results under different number of people in a group. The probability p_1 indicates that obtained due to Equation 4, p_2 that due to Equation 10 when taking into consideration Equation 5, we obtain probability p_3 as

$$p_3 \approx p_2 - \sum_{1=i<}^{r-2}\sum_{j}^{r-1} 1.1.1...(\frac{i}{365}).(\frac{j}{365}) \text{ and } p_4 \text{ as}$$

$$p_4 \approx p_3 + \sum_{1=i<}^{r-3}\sum_{j}^{r-2}\sum_{k}^{r-1} 1.1.1...(\frac{i}{365}).(\frac{j}{365}).(\frac{k}{365})$$

while the differences (errors) of these p values from p_1 values are $d_1 = p_2 - p_1$, $d_2 = p_3 - p_1$, $d_3 = p_4 - p_1$.

The models were coded in Pascal programming language in order to generate Table 1. The data in Table 1 were used to obtain Figs. 1 and 2 using Microsoft Excel. The program is presented as

```
Begin
 clrscr;
  writeln;
        r:=2;
     writeln( 'r     p1       p2       p3       p4       d1       d2       d3');
     writeln('-----------------------------------------------------------------');
     Repeat
        c:=1; w:=0; s:=0; t:=0;
        for i:=365-r+1 to 365 do
          c:=c*i/365;
         p1:=1-c;
         p2:=r*(r-1)/730;
         for j:=2 to r-1 do
           for i:=1 to j-1 do
             s:=s+(i/365*j/365);
           for k:=3 to r-1 do
             for j:=2 to k-1 do
               for i:=1 to j-1 do
                 t:=t+(i/365*j/365*k/365);
            for v:=4 to r-1 do
              for k:=3 to v-1 do
                for j:=2 to k-1 do
                  for i:=1 to j-1 do
                    b:=b+(i/365*j/365*k/365*v/365);
          p3:=p2-s; p4:=p3+t; d1:=p2-p1; d2:=p1-p3; d3:=p4-p1;
       writeln(r:6,'  ',p1:6:6,'  ',p2:6:6,'  ',p3:6:6,'  ',p4:6:6,'  ',d1:6:6,'            ',d2:6:6,'  ',d3:6:6);
    r:=r+1;
  until r=m;
end.
```

From Table 1 and Fig. 2, if we do not want to incur an error (difference between p_1 with others) of not more than .05, we will not wish to use p_2 as explained above when the number in a group does exceed 16. While for p_3 the number should not exceed 23. For p_4 the number should not exceed 29. The p_2 estimate fails to estimate well when the number in a group becomes large as for example 28 we see that the probability exceeds 1, which is outrageous. This is so because the product $r(r\text{-}1) = (28)(27)$ exceeds 730. p_3 never exceeds 1, but that it starts to decline when the number in a group is more than 28 which fails to approximate p_1. p_4 is better of the estimates, though it also exceeds probability of 1 when the number in a group exceeds 35. The d_3s' are quite smaller in value depicting how good the estimate p_4 of p_1. The d_1 successively increases to more than 1 when the number in a group goes to beyond 40 due to the reason given of p_2 above.

Fig.1 PROBABILITY OF AT LEAST TWO PEOPLE SHARING SAME BIRTHDAY

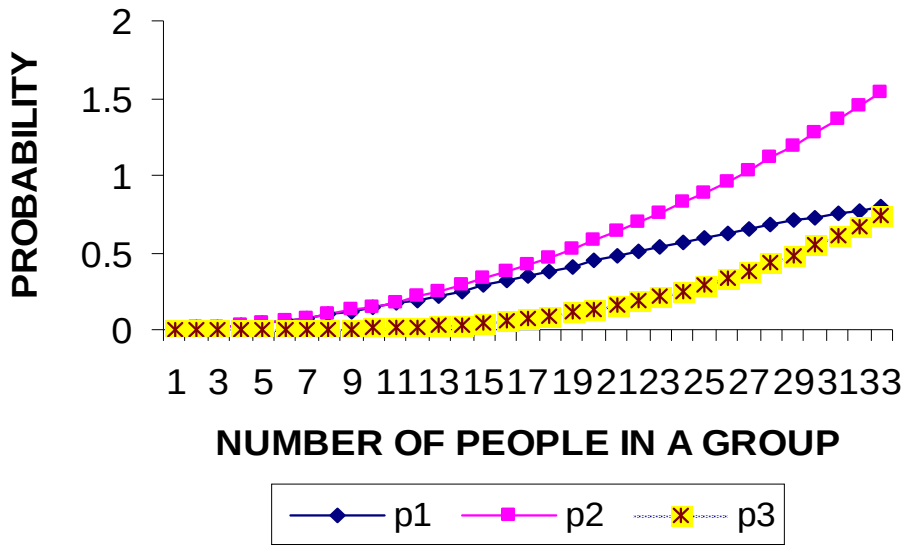

Fig. 2 PROBABILITY OF AT LEAST TWO PEOPLE SHARING A BIRTHDAY

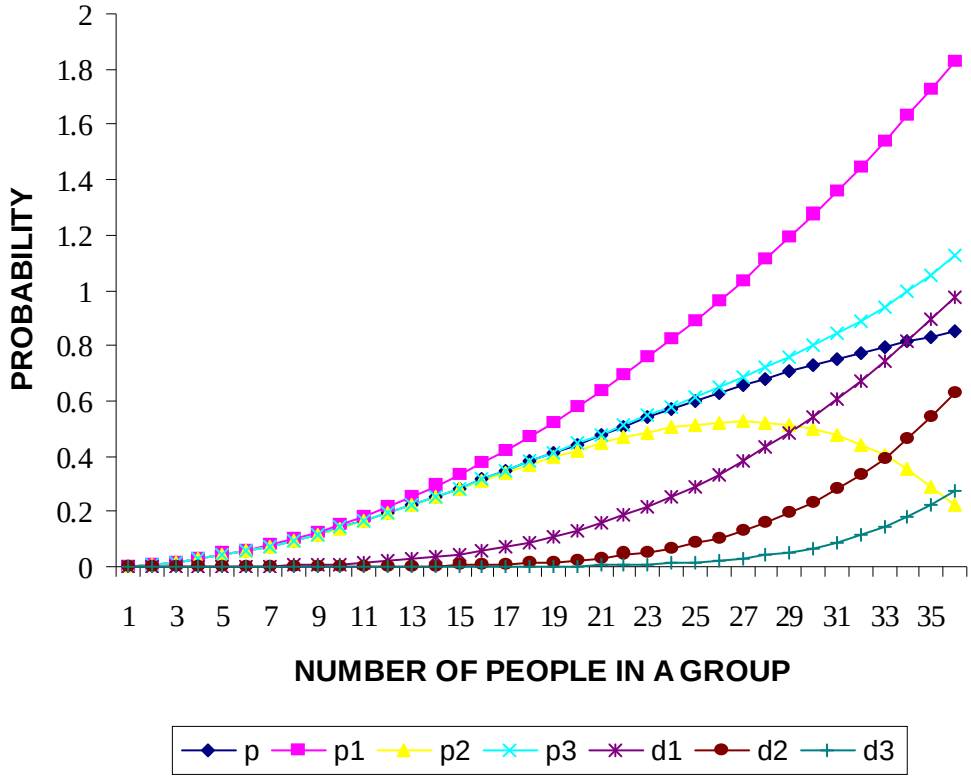

TABLE 1. PROBABILITIES OF AT LEAST TWO PEOPLE IN A GROUP OF SIZE *N* SHARING SAME BIRTHDAY UNDER VARIOUS SITUATIONS ALONG WITH THEIR RESPECTIVE DEVIATIONS.

r	p1	p2	p3	p4	d1	d2	d3
2	0.002740	0.002740	0.002740	0.002740	0	0	0
3	0.008204	0.008219	0.008204	0.008204	0.000015	0	0
4	0.016356	0.016438	0.016356	0.016356	0.000082	0	0
5	0.027136	0.027397	0.027135	0.027136	0.000262	0.000001	0
6	0.040462	0.041096	0.040458	0.040462	0.000633	0.000005	0
7	0.056236	0.057534	0.056221	0.056236	0.001299	0.000015	0
8	0.074335	0.076712	0.074295	0.074336	0.002377	0.000040	0
9	0.094624	0.098630	0.094532	0.094625	0.004006	0.000092	0.000001
10	0.116948	0.123288	0.116757	0.116952	0.006339	0.000191	0.000004
11	0.141141	0.150685	0.140777	0.141150	0.009544	0.000364	0.000009
12	0.167025	0.180822	0.166373	0.167045	0.013797	0.000652	0.000020
13	0.194410	0.213699	0.193305	0.194451	0.019288	0.001106	0.000041
14	0.223103	0.249315	0.221310	0.223183	0.026213	0.001793	0.000081
15	0.252901	0.287671	0.250103	0.253051	0.034770	0.002798	0.000149
16	0.283604	0.328767	0.279377	0.283868	0.045163	0.004227	0.000264
17	0.315008	0.372603	0.308801	0.315457	0.057595	0.006207	0.000450
18	0.346911	0.419178	0.338022	0.347650	0.072267	0.008889	0.000739
19	0.379119	0.468493	0.366665	0.380296	0.089375	0.012453	0.001177
20	0.411438	0.520548	0.394333	0.413264	0.109110	0.017105	0.001825
21	0.443688	0.575342	0.420604	0.446451	0.131654	0.023084	0.002763
22	0.475695	0.632877	0.445037	0.479786	0.157181	0.030659	0.004091
23	0.507297	0.693151	0.467165	0.513236	0.185853	0.040133	0.005939
24	0.538344	0.756164	0.486500	0.546812	0.217820	0.051844	0.008468
25	0.568700	0.821918	0.502533	0.580576	0.253218	0.066166	0.011876
26	0.598241	0.890411	0.514731	0.614649	0.292170	0.083510	0.016408
27	0.626859	0.961644	0.522537	0.649216	0.334785	0.104322	0.022357
28	0.654461	1.035616	0.525374	0.684536	0.381155	0.129087	0.030074
29	0.680969	1.112329	0.522642	0.720945	0.431360	0.158326	0.039977
30	0.706316	1.191781	0.513717	0.758872	0.485465	0.192599	0.052556
31	0.730455	1.273973	0.497955	0.798841	0.543518	0.232500	0.068386
32	0.753348	1.358904	0.474686	0.841480	0.605557	0.278662	0.088133
33	0.774972	1.446575	0.443220	0.887535	0.671603	0.331752	0.112563
34	0.795317	1.536986	0.402845	0.937874	0.741669	0.392472	0.142557
35	0.814383	1.630137	0.352824	0.993500	0.815754	0.461559	0.179116
36	0.832182	1.726027	0.292400	1.055557	0.893845	0.539782	0.223375
37	0.848734	1.824658	0.220792	1.125348	0.975924	0.627942	0.276614

From the above work, it is realized that the fewer the people in a group the greater the chance that they are born on different days of the year. When number of people in a group is as small as 23, the probability that at least two of them share a birthday is greater than ½. This shows that we do not need have half the days of the year to attend the probability ½. The approximation model gives good estimates when the number in a group is small, but requires more terms of the expansion when the number in a group becomes large.

REFERENCES

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, New York.

Levine, G. & Burke, C. J. (1972). *Mathematical Model Techniques For Learning Theories*. Academic Press, New York.

Ross, S. (1976). *A First Course in Probability*. Macmillan Publishing Co. New York.

Snell, J. L. (1987). *Introduction to Probability*. Random House/ Birkhauser. Mathematics Series New York.