TRANSACTIONS OF SOCIETY OF ACTUARIES
1959 VOL. 11 NO. 29AB

ON COMPUTING THE PROBABILITY THAT EXACTLY k OF n INDEPENDENT EVENTS WILL OCCUR

ROBERT P. WHITE* AND T. N. E. GREVILLE

INTRODUCTION

IN THE valuation of estates it is sometimes necessary to compute the probability that exactly k (or at least k) out of n designated lives will survive a given number of years. It is the purpose of this note to explain a method of computation which may often be more expeditious than the methods in common use, particularly when the number of lives is large.

For purposes of mathematical analysis, we may as well generalize the problem to refer to n independent events of any kind, with respective probabilities of occurrence of $p_1, p_2, \ldots, p_n$ (and respective probabilities of nonoccurrence of $q_1, q_2, \ldots, q_n$, where $q_i = 1 - p_i$). The probability that exactly k of the n events will occur is the sum of $C(n, k)$ terms, each of which is the product of k of the p's and the $n - k$ q's corresponding to the remaining p's. When n is large, the amount of computation involved can be formidable. For example, the probability that exactly 8 out of 28 events will occur is the sum of $C(28, 8) = 3{,}108{,}105$ such products.

If we make use of Waring's theorem [**1**, p. 74][1] (also called the "method of inclusion and exclusion"), which is commonly known among actuaries as the "Z method" [**2**, p. 189], the probability that exactly k of the n events will occur is given by

$$B_k - (k+1)B_{k+1} + \ldots + (-1)^{r-k} C(r, k) B_r + \ldots + (-1)^{n-k} C(n, k) B_n\,,$$

where B_r is the sum, for all possible choices of r events out of the n, of the probabilities that r specified events will happen, irrespective of whether the other $n - r$ events occur. Under this procedure the number of products required is substantially greater than for the previous method. However, the number of multiplications to be performed is somewhat smaller, if the individual products are recorded at each stage and account is taken of the fact that any product of $r + 1$ p's can be obtained by a single mul-

* Robert P. White, not a member of the Society, is employed in the Internal Revenue Service.

[1] Boldface figures within square brackets refer to the bibliography at the end of the paper.

tiplication from some product of r p's. The necessary amount of recording is, of course, substantially increased.[2]

It is the purpose of the present note to describe two further methods of computing this probability: one based on the elementary rules for combination of probabilities, and the other on the use of a generating function and the relations between the roots and coefficients of a polynomial. In the application to problems in life contingencies, frequently much labor is saved by these methods as compared with those referred to earlier.

METHOD OF COMPOSITION OF PROBABILITIES[3]

Let ${}_mP_{[r]}$ denote the probability that exactly r out of the first m events occur (irrespective of the outcome of the remaining $n - m$ events). Now let us analyze this probability by considering the first m events as made up of two groups: the first $m - 1$ events and the mth event. Then there are two ways in which we can have exactly r of the first m events occurring, as follows:

(i) exactly $r - 1$ of the first $m - 1$ events occur, and the mth event occurs,

(ii) exactly r of the first $m - 1$ events occur, and the mth event does not occur.

This analysis gives at once the equation:

$$ {}_mP_{[r]} = p_m \cdot {}_{m-1}P_{[r-1]} + q_m \cdot {}_{m-1}P_{[r]} \, , \tag{1} $$

where, of course, ${}_lP_{[s]}$ is to be interpreted as 0 for $s < 0$ or for $s > l$. By the use of this equation a table of values of ${}_mP_{[r]}$ is easily built up. A convenient arrangement is illustrated in Table 1, where the first two columns show the values of p_m and q_m and the remaining columns give values of ${}_mP_{[r]}$, with the row corresponding to the value of m, and the column to the value of r. With this arrangement the complete table of values of ${}_mP_{[r]}$ for $m, r = 1, 2, \ldots, n$ $(r \leq m)$ forms a right triangle with the right angle in the lower left corner.

If the value of ${}_nP_{[k]}$ is required for only a single value of k, only those values of ${}_mP_{[r]}$ need be computed which lie in a parallelogram $k + 1$ columns wide and $n - k + 1$ lines deep having one of its oblique sides along the hypotenuse of the right triangle. The entire calculation then requires $n - 2$ simple multiplications and $k(n - k)$ operations of the type $pr + qs$. In the example previously referred to (computation of ${}_{28}P_{[8]}$) only 186 values must be computed, involving a total of 346 multiplications.

[2] In comparison with other methods Waring's theorem has the advantage that it applies even when the events are not independent [1, p. 86].

[3] Our discussion of this method is the result of a suggestion made by Mr. David W Bennett, a graduate student at Columbia University.

TABLE 1

CALCULATION OF ${}_{10}P_{[k]}$ FOR $k = 0, 1, 2, 3, 4$

Event No.	Probability of Occurrence p_i	Probability of Nonoccurrence q_i	${}_iP_{[0]}$	${}_iP_{[1]}$	${}_iP_{[2]}$	${}_iP_{[3]}$	${}_iP_{[4]}$
1........	.91245 53172	.08754 46828	.08754 46828	.91245 53172	....	....	....
2........	.85480 30981	.14519 69019	.01271 12167	.20731 91513	.77996 96320	....	....
3........	.25589 79743	.74410 20257	.00945 84421	.15751 93751	.63342 95340	.19959 26488	....
4........	.13045 55591	.86954 44409	.00822 45358	.13820 40033	.57134 44081	.25618 90822	.02603 79706
5........	.95751 29534	.04248 70466	.00034 94362	.01374 69795	.15660 68599	.55795 43891	.24641 06412
6........	.94122 10435	.05877 89565	.00002 05395	.00113 69298	.02214 41342	.18019 76489	.53964 21727
7........	.89465 82204	.10534 17796	.00000 21637	.00013 81420	.00334 98661	.03879 37727	.21806 21747
8........	.46195 16413	.53804 83587	.00000 11642	.00007 53266	.00186 62049	.02242 04019	.13524 88422
9........	.14654 95518	.85345 04482	.00000 09936	.00006 44581	.00160 37525	.01940 81935	.11871 38848
10........	.05503 53085	.94496 46915	.00000 09389	.00006 09653	.00151 90370	.01842 83206	.11324 85654

If the values of ${}_nP_k$, the probability that *at least* k of the n events occur, are desired, no simple relationship corresponding to (1) is available, and it is necessary to compute ${}_nP_{[0]}, {}_nP_{[1]}, \ldots, {}_nP_{[k-1]}$, and then make use of the relation

$$ {}_nP_r = {}_nP_{r-1} - {}_nP_{[r-1]}\,, $$

where of course ${}_nP_0 = 1$. We must then compute and record *all* values in the first $k + 1$ columns of the table of values of ${}_mP_{[r]}$ (enlarging the parallelogram above to a trapezoid). The entire calculation then involves $n + k - 2$ simple multiplications and $\frac{1}{2}k(2n - k - 1)$ operations of the type $pr + qs$.

GENERATING FUNCTION METHOD

Fréchet and Bizley [4, pp. 11–15; 1, p. 62] have pointed out that $P_{[k]}$ (omitting the subscript n) is the coefficient of x^k in the expansion of

$$ \phi(x) = (p_1x + q_1)(p_2x + q_2)\ldots(p_nx + q_n)\,. \tag{2} $$

Thus the function $\phi(x)$ may be described as a generating function for the probabilities $P_{[k]}$. Bizley states that "where the p_i are not all equal, . . . [this] product . . . is often useful and in numerical examples may provide a quicker solution than . . . Waring's theorem."

In fact, the method of composition of probabilities previously described is easily deduced from the right member of (2). However, a different approach is suggested by well known relations between the roots and coefficients of a polynomial. If s_r denotes the sum of the rth powers of the roots of the equation

$$ x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n = 0\,, \tag{3} $$

then, according to a formula attributed to Newton [3, p. 437],

$$ s_k + a_1s_{k-1} + a_2s_{k-2} + \ldots + a_{k-1}s_1 + ka_k = 0\,, \tag{4} $$

for $k = 1, 2, \ldots, n$.

Now, it is clear from general considerations as well as from (2) that

$$ P_{[n]} = Q_{[0]} = p_1p_2\ldots p_n\,, \tag{5} $$

where $Q_{[k]}$ is used to denote the probability that exactly k of the n events fail. Moreover, it is evident that equation (2) will be in the form (3) if it is divided by the quantity (5). It is obvious also that the roots of (2) are $-q_1/p_1, -q_2/p_2, \ldots, -q_n/p_n$. Thus, if we define

$$ T_r = (-1)^r s_r = (q_1/p_1)^r + (q_2/p_2)^r + \ldots + (q_n/p_n)^r $$

and note that $a_r = Q_{[r]}/Q_{[0]}$, then (4) gives

$$Q_{[k]} = \frac{1}{k}\,[T_1Q_{[k-1]} - T_2Q_{[k-2]} + T_3Q_{[k-3]} - \ldots + (-1)^{k-1}T_kQ_{[0]}]\,. \quad (6)$$

It is perhaps generally more convenient to interchange p's and q's, P's and Q's, and to write (6) in the form

$$P_{[k]} = \frac{1}{k}\,[S_1P_{[k-1]} - S_2P_{[k-2]} + S_3P_{[k-3]} - \ldots + (-1)^{k-1}S_kP_{[0]}]\,, \quad (7)$$

$$S_r = (p_1/q_1)^r + (p_2/q_2)^r + \ldots + (p_n/q_n)^r\,. \quad (8)$$

Formulas (7) and (8), together with the obvious relation

$$P_{[0]} = q_1q_2 \ldots q_n\,, \quad (9)$$

provide a method of computing successively $P_{[0]}, P_{[1]}, \ldots P_{[k]}$.

COMPARISON OF THE METHODS

The calculation of $P_{[k]}$ by the second method requires n divisions and $n(k-1)$ multiplications to obtain the powers of p_i/q_i, followed by the summing of these powers, one application of formula (9) and k applications of formula (7). This makes a total of $k(n+2)+1$ values to be computed and recorded, as compared with $(k+1)(n-k+1)-3$ under the method previously described. On this basis alone, the method last described would appear to be more advantageous only when $n > k^2 + 2k + 3$. This criterion would limit the use of the second method to cases in which k is quite small in comparison to n.

This does not take into account, however, that some of the values to be recorded require only a simple multiplication, while others, such as application of formula (7), involve a fairly complex sequence of arithmetical operations. Thus another possible criterion would be the total number of individual multiplications and divisions required. The first method requires a total of $2k(n-k)+n-2$ multiplications; for the second, the combined total of multiplications and divisions is $(n + \frac{1}{2}k + 1)(k-1) + 2n + k - 1$. On this basis the second method is preferable for $n > \frac{1}{2}(5k + 3)$. In the example of calculating ${}_{28}P_{[8]}$ there are 186 values to be computed and recorded under the first method; 241 under the second. The number of individual multiplications and divisions required in this example is 346 under the first method and 286 under the second.

The second method frequently involves less actual labor, but it usually requires more recording, and it is less straightforward, since it consists of several distinct steps, while the first method involves merely repeated application of one simple formula. Another characteristic of the second method which some users may consider a disadvantage is that the different quantities which will need to be computed usually differ widely in magnitude, and yet all must be carried out to several significant figures in order to avoid excessive accumulation of rounding error. Thus it becomes necessary to record the data in scaled form, as illustrated in Tables 2 and 3. In other words, the decimal point is maintained in a fixed position relative to the first significant digit of the number, and the resulting number is multiplied by an appropriate power of 10 to bring it to the correct order of magnitude. This method of handling numbers is in common use among physicists, engineers, and electronic computer users, and does not in itself really involve more work. However, it might in some situations constitute a psychological barrier to the use of the second method. If a problem of the

TABLE 2

CALCULATION OF POWERS OF p_i/q_i

Event No.	Probability p_i	p_i/q_i	$(p_i/q_i)^2$	$(p_i/q_i)^3$	$(p_i/q_i)^4$
1......	.91245 53172	10.4227 3829	108.6334 735	1,132.25826	11,801.2316
2......	.85480 30981	5.8871 9929	34.6591 155	204.04512	1,201.2543
3......	.25589 79743	.3439 0173	.1182 684	.04067	.0140
4......	.13045 55591	.1500 2748	.0225 082	.00338	.0005
5......	.95751 29534	22.5365 8538	507.8976 806	11,446.27944	257,960.0539
6......	.94122 10435	16.0128 9134	256.4126 891	4,105.90853	65,747.4671
7......	.89465 82204	8.4929 0969	72.1295 150	612.58946	5,202.6669
8......	.46195 16413	.8585 6900	.7371 407	.63289	.5434
9......	.14654 95518	.1717 1419	.0294 858	.00506	.0009
10......	.05503 53085	.0582 4060	.0033 920	.00020	.0000
Total.	....	64.9347 7699	980.6432 688	17,501.76301	341,913.2326

type under consideration were to be programmed for an electronic computer, it would probably be advisable to avoid the use of the second method unless the computer has floating-point hardware.

The probability that exactly k of the n events will *fail* can be obtained in a comparable manner by substituting p_i for q_i and q_i for p_i in the above equations. The probability that exactly k of the events will occur being equal to the probability that exactly $n - k$ of the events will fail, labor will sometimes be saved under the second method by computing the latter probability rather than the former, depending on the magnitude of the ratios p_i/q_i and whether k is closer to n than to 0. Under the first method there is no saving in labor (unless the probability that at least, or at most, k lives will fail is required), but, when k is closer to n than to 0, a more compact arrangement of the table is secured by considering failure rather than occurrence of the events.

PROPERTIES OF THE GENERATING FUNCTION

It is interesting to note in passing that both Waring's formula and the formula (6) of Rasor and Myers (*TSA* IV, 128) are easily derived from the generating function (2).

If we write $u = 1 - x$, (2) can be put in the form

$$\phi(x) = \psi(u) = (1 - p_1 u)(1 - p_2 u) \ldots (1 - p_n u) .$$

It is then easily verified that

$$\psi(u) = 1 - B_1 u + B_2 u^2 - \ldots + (-1)^n B_n u^n , \qquad (10)$$

where B_r denotes, as before, the sum, for all selections of r events out of the n, of the probabilities that r specified events occur, irrespective of the

TABLE 3

CALCULATION OF $P_{[r]}$

r	s_r	$P_{[r]}$
0	....	$.93887\ 04315 \times 10^{-6}$
1	$.64934\ 77699 \times 10^{2}$	$.60965\ 34209 \times 10^{-4}$
2	$.98064\ 32688 \times 10^{3}$	$.15190\ 36962 \times 10^{-2}$
3	$.17501\ 76301 \times 10^{5}$	$.18428\ 32060 \times 10^{-1}$
4	$.34191\ 32326 \times 10^{6}$	$.11324\ 85652$

outcome of the remaining $n - r$ events. Differentiating both (2) and (10) k times with respect to x and then setting $x = 0$ (and therefore $u = 1$) gives Waring's formula.

On the other hand, integration of (2) and (10) with respect to x between the limits 0 and 1 gives

$$P_{[0]} + \tfrac{1}{2}P_{[1]} + \tfrac{1}{3}P_{[2]} + \ldots + \frac{1}{n+1} P_{[n]}$$
$$= 1 - \tfrac{1}{2}B_1 + \tfrac{1}{3}B_2 - \ldots + \frac{(-1)^n}{n+1} B_n ,$$

from which Rasor and Myers' formula follows at once. Approximate integration for the right member of (2), between the limits 0 and 1 for x, may often be a simple way of obtaining a good approximation to this value.

NUMERICAL EXAMPLE

Given 10 lives consisting of male lives aged 44, 50, 77, and 82 and female lives aged 42, 46, 53, 73, 83, and 88, let it be required to find the probability that at least 5 survive 10 years on the basis of the 1949–51 United States life tables for white males and white females. Table 1 shows the computation of the values of $P_{[k]}(k = 0, 1, \ldots, 4)$ under the method of composition of probabilities. (As previously explained, the lower left portion of the table, exclusive of the first two columns, could have been omitted if only P_4 had been required.) Tables 2 and 3 show the computation of the same five quantities by the generating function method. The work has been carried out to the full capacity of the calculating machine, and numbers are recorded in scaled form, as previously explained, in order to minimize rounding error. In Table 3, $P_{[0]}$ was obtained by (9) and then $P_{[1]}$, $P_{[2]}$, $P_{[3]}$, and $P_{[4]}$ were computed by the formulas

$$P_{[1]} = S_1 P_{[0]} ,$$

$$P_{[2]} = \tfrac{1}{2}(S_1 P_{[1]} - S_2 P_{[0]}) ,$$

$$P_{[3]} = \tfrac{1}{3}(S_1 P_{[2]} - S_2 P_{[1]} + S_3 P_{[0]}) ,$$

$$P_{[4]} = \tfrac{1}{4}(S_1 P_{[3]} - S_2 P_{[2]} + S_3 P_{[1]} - S_4 P_{[0]}) ,$$

all derived from (7).

TABLE 4

VALUES OF $P_{[r]}$ AND P_r

r	$P_{[r]}$	P_r
0	.00000	1.00000
1	.00006	1.00000
2	.00152	.99994
3	.01843	.99842
4	.11325	.97999
5	....	.86674

Table 4 shows the calculation of values of P_k (under either method) by successive application of the relation ${}_nP_r = {}_nP_{r-1} - {}_nP_{[r-1]}$.

BIBLIOGRAPHY

1. M. T. L. BIZLEY. ***Probability, an Intermediate Text-Book.*** Published for the Institute of Actuaries and Faculty of Actuaries by the Cambridge University Press, 1957.
2. C. W. JORDAN, JR. ***Society of Actuaries Textbook on Life Contingencies.*** Society of Actuaries, Chicago, 1952.
3. G. CHRYSTAL. ***Algebra, an Elementary Textbook,*** Vol. 1. Chelsea Publishing Co., New York, 1952.
4. M. FRÉCHET. ***Les Probabilités Associées à un Système d'Evénements Compatibles et Dépendants, Première Partie, Evénements en Nombre Fini Fixe.*** ("Actualités Scientifiques et Industrielles," No. 859.) Hermann & Cie., Paris, 1940.