THE RELATION BETWEEN THE DISTRIBUTION OF SICKNESS AND THE EFFECT OF DUPLICATES ON THE DISTRIBUTION OF DEATHS

By R. E. BEARD, M.B.E., F.I.A.

Assistant General Manager, Pearl Assurance Company, Ltd.

AND WILFRED PERKS, F.I.A.

Joint Actuary, Pearl Assurance Company, Ltd.

1. In *J.S.S.* Vol. VII, Pt. 1, pp. 23–8 (1947), one of the present writers (R.E.B.) discusses the standard deviation of the distribution of sickness. In *Skandinavisk Aktuarietidskrift*, 1947, Häft 1–2, pp. 18–43, H. L. Seal discusses the effect of duplicates on the distribution of deaths. The purpose of this note is to show the theoretical relation between the two problems, to bring out the need in theoretical problems of this kind to specify clearly the sampling process contemplated in the mathematical solution, to remove certain weaknesses in Seal's analysis and to provide a critical analysis of the use of the Poisson law in the theory of actuarial statistics.

2. In the problems of sickness and of duplicates it is assumed that we are concerned with sampling from a 'large' population distributed in the following forms:

Sickness

p = proportion of lives not claiming during the 'risk period',

q = proportion claiming, π_t = proportion of those claiming who are sick for exactly duration t in the 'risk period'. It is assumed that a person can make only one claim in the 'risk period', i.e. that all attacks are linked up under an off-period rule. Sickness of some specified kind (e.g. 'Second three months') is contemplated and duration t is the period of sickness of this kind in the period of exposure. Thus, if in the period of exposure a man is sick from the 15th to the 19th weeks in the M.U. sense, t is 4 if the unit of time is taken as a week.

If the 'risk period' is taken as the unit of time, the proportion of the 'risk period' during which there is sickness is $q \sum_1 t\pi_t$ on the average.

Duplicates

p = proportion of lives surviving the 'risk period', q = proportion dying,

π_t = proportion of those living and of those dying (or of both together) who hold t policies, i.e. mortality is assumed to be independent of the number of policies.

The proportion of policies becoming claims is $q \sum_1 t\pi_t$.

Thus for both problems we have a probability distribution in the form

$$f(0)=p, \quad f(1)=q\pi_1, \quad f(2)=q\pi_2, \quad \ldots, \quad f(t)=q\pi_t, \quad \ldots,$$

where $p+\sum_1 q\pi_t = 1$, and the first moment of this distribution, viz. $q \sum_1 t\pi_t$,

represents the sickness experienced or policies becoming claims as the case may be. The second moment is $q \sum_1 t^2\pi_t$, and the variance is, therefore,

$$q\{\sum_1 t^2\pi_t - q(\sum_1 t\pi_t)^2\}.$$

3. Thus, for samples of N lives selected independently at random the expected sickness and the expected claims (i.e. policies becoming claims) are both represented by $Nq \sum_1 t\pi_t$ and the variances by $Nq\{\sum_1 t^2\pi_t - q(\sum_1 t\pi_t)^2\}$, because of the additive property of variances (and of higher cumulants) in the case of independent sampling. The variance can be put in the form

$$Nq \sum_1 t\pi_t \left\{ \frac{\sum_1 t^2\pi_t}{\sum_1 t\pi_t} - q \sum_1 t\pi_t \right\}. \qquad (1)$$

4. In the sickness problem the term $Nq \sum_1 t\pi_t$ represents the expected duration of sickness claims, and the exact expression for the variance can be compared with the approximate expression given by R. E. Beard, *J.S.S.* Vol. VII, p. 26, in which the second term inside the brackets, viz. $q \sum_1 t\pi_t$, is omitted. This term is clearly small in relation to the first term, and its omission would not greatly affect the variance. The chief merit of the approximate form is that, by the substitution of the actual sickness claims for the expected claims and the use of the ratio of the second moment to the first moment of the actual claims instead of $\sum_1 t^2\pi_t / \sum_1 t\pi_t$, an approximate variance can be calculated from the actual claims alone, without reference to the exposed to risk. The development of Beard's analysis using the binomial law as indicated at the foot of p. 23 in *J.S.S.* Vol. VII would, of course, lead to the expression (1) above.

5. The basis of the foregoing formula is that the sampling is for a fixed number of lives, and that the sampling process is not otherwise restricted. This seems to be appropriate for the sickness problem and, if the data are supplied in the form contemplated, little more need be said. However, published sickness data are usually not in this form, and the more complicated problem arising when the data are in the form of sickness rates for each week of sickness has been treated by L. E. Coward (*J.I.A.* Vol. LXXV, p. 12). This variation of the problem which was touched upon by Beard does not, however, involve any difference so far as the sampling process is concerned.

6. In the case of duplicate policies Seal clearly has doubts about the suitability of assuming a sampling process in which the number of lives is kept constant from sample to sample so that the number of policies fluctuates, and he has posed the problem from the point of view of keeping the number of policies constant, thus permitting the number of lives to fluctuate from sample to sample. In the former case the sampling process assumed is obvious.

7. In the case posed by Seal, however, the sampling process assumed is by no means so obvious or so realistic. It is of no use to take direct samples of E policies. If this were contemplated, the samples would be quite unrepresentative of the universe of policies; indeed, it would be an unusual event for such

a sample of moderate size to contain any duplicates at all. The essential unit for sampling is the life. The sampling process might be assumed to take the form of the random selection of lives until the associated policies equal or exceed E;* only those samples in which the number of associated policies is exactly E would be retained, the rest being discarded. Or we might select a number of lives at random and then observe whether they hold exactly E policies. For our part we cannot see that such processes represent realistic positions to take in relation to the practical problems of the distribution of claims by death when duplicate policies are included in the experience. It is significant that in obtaining his own sample Seal selected 2000 lives and recorded the distribution of policies amongst them. He rounded off the number of lives and not the number of policies! (We have to bear in mind that a life office obtains its business by its representatives selecting lives). Incidentally, Seal does not tell us whether the sample includes all duplicates or whether the rules for elimination of concurrent duplicates applicable to the data of the A 1924–29 table were used. Further, the duplicate distributions applicable to a particular office may be quite unrepresentative of that either of the offices as a whole or of the particular mixture of offices included in the A 1924–29 experience.

8. Perhaps the most suitable sampling process to contemplate in relation to the problem of duplicates is a process of stratified sampling, i.e. the random selection of $\pi_1 \mathrm{N}$ lives with only one policy, $\pi_2 \mathrm{N}$ lives with exactly two policies and so on. In this process π_t is no longer treated as a source of random variation; we ask ourselves how the variance is affected by weighting the deaths by the number of policies held. The variance on this basis is given by Seal in an appendix to his paper on Graduation Tests, *J.I.A.* Vol. LXXI, p. 40, and may be written in our notation as $\mathrm{N}pq\Sigma t^2\pi_t$, where the π_t are now the actual proportions of the lives with 1, 2, 3, ... policies respectively. This variance has long been appreciated in the form $\Sigma \mathrm{S}^2 pq$, where S is the sum assured on the life, i.e. the 'weight'.

9. Incidentally, in deriving this result Seal assumes, in addition to the essential condition of independence in the sampling, that the lives are homogeneous. This latter condition is more restrictive than is necessary; all that is necessary, in addition to independent sampling, is that the mortality should be independent of the number of policies on the life. Even this restriction can be removed from (1).

10. If we write n_t as the number of lives who die with t policies on their lives and l_t as the number who survive with t policies on their lives the following cases arise:

(*a*) For simple sampling of lives, without stratification, the relation

$$\sum_1 n_t + \sum_1 l_t = \mathrm{N}$$

must hold and the variance given earlier may be written in the form

$$\mathrm{N}qm_2 - \mathrm{N}q^2 m_1^2, \qquad (2)$$

where m_r is the rth moment about zero.

* A full discussion of similar practical problems is given in Yates's paper in *J.R.S.S.* Vol. CIX, p. 12. The remarks by Kendall and by Anscombe and Quenouille should be noted, particularly the difficulty arising when the population is J-shaped.

(*b*) For stratified sampling the relations $n_t + l_t = N\pi_t (t = 1, 2, \ldots)$ all hold separately, and hence $\sum_1 n_t + \sum_1 l_t = N$ also holds. In this case the variance may be written in the form

$$Npqm_2 = Nqm_2 - Nq^2(\mu_2 + m_1^2) = Nqm_2 - Nq^2 m_1^2 - Nq^2\mu_2, \qquad (3)$$

where μ_r is the rth moment about the mean.

*(*c*) In the case considered by Seal the relation

$$\sum_1 tn_t + \sum_1 tl_t = E$$

holds, and the variance for large E in this case, based on a binomial distribution, is

$$pq\left(E\frac{\sigma_2}{\sigma_1} + \frac{\sigma_2^2}{\sigma_1^2} - \frac{\sigma_3}{\sigma_1}\right) \quad \text{(see appendix A).}$$

Noting that $\sigma_r = m_r$ and that $E = Nm_1$, this may be written

$$Nqm_2 - Nq^2 m_1^2 - Nq^2\mu_2 - pq\left(\frac{m_3}{m_1} - \frac{m_2^2}{m_1}\right). \qquad (4)$$

*(*d*) A sampling process can be defined in which $\Sigma n_t + \Sigma l_t \neq N$ by considering a set of samples of N lives and selecting the π_1 group from the first sample, the π_2 group from the second and so on. The variance in this case is

$$Nqm_2 - Nq^2 m_1^2 + 2Nq^2 \Sigma jk\pi_j\pi_k \quad (j < k). \qquad (5)$$

If each of the π_t groups is treated as a Poisson variable and the sampling is independent, the variance (see R. E. Beard, *J.S.S.* Vol. VII, p. 25) is Nqm_2. It will be seen that, if the number of groups is large and each value of π_t tends to zero, the second and third expressions in (5) tend to counterbalance.

Case (*d*) represents the variation with no constraints, and the other cases reflect the reduction in variance due to the imposition of different constraints. Thus, in case (*a*) the constraint $\sum_1 n_t + \sum_1 l_t = N$ is imposed and the variance is reduced by the last term in (5).

The last term in (3) shows the further reduction in variance occasioned by the additional constraints of stratified sampling. This term is quite small for the values of q that arise in practice.

11. It is interesting that the case studied by Seal shows a variance slightly less than that of the stratified sampling case, although only one constraint ($\sum_1 tn_t + \sum_1 tl_t = E$) is imposed. The last term in (4) becomes less and less important as E increases, until, in the limit, Seal's case becomes identical with the 'stratified' case. The expression (4) is, of course, appropriate only for large E. The reason why Seal's case approximates to the stratified case with its m constraints appears to be that the condition $\sum_1 tn_t + \sum_1 tl_t = E$, though apparently a single constraint, is closely associated with the problem of selection of integers from the set $1, 2, \ldots, n$, repetitions being allowed, such that their sum = E. It seems that each integer gives rise, in effect, to an element of constraint, so that the result approximates to that obtained by stratified sampling. For small values of E only very restricted combinations may be possible to obtain a sample

* The derivation of the variances in cases (*c*) and (*d*) is given in the Appendix by R. E. Beard.

complying with Seal's condition because of the restriction to integers. In the stratified case this point does not arise because it is implicitly assumed that the sample proportions are, in fact, those of the universe.

12. It is worth noting that a sampling process could be defined in which both of the conditions $\sum_1 n_t + \sum_1 l_t = \mathrm{N}$ and $\sum_1 t n_t + \sum_1 t l_t = \mathrm{E}$ apply. If no restriction is placed on the relative values of N and E some light is thrown upon the nature of the relationship of Seal's case to the others. Thus if $\mathrm{N} = \mathrm{E}$ all the cases must be from the π_1 group (i.e. single lives with one policy only) and the variance becomes $pq\mathrm{E}$. At the other extreme, N could be made equal to E/m, where m is the maximum number of policies per life, so that all cases have to be selected from the π_m group. In this case the variance becomes $pq\mathrm{N}m^2 = pq\mathrm{E}m$. For values of N between these limits the variance will lie between the extremes $pq\mathrm{E}$ and $pq\mathrm{E}m$, but a smooth progression would appear to be unlikely. For the samples to be representative some restriction would have to be placed on N, leading to what has been called 'balanced sampling' (see Yates, *loc. cit.*), and some such relationship as $\mathrm{E} = \mathrm{N} \sum_1 t\pi_t$ would appear to be necessary.

13. A general solution of this last case has not yet been found (see Appendix) but the variance in the particular case when $m = 2$ has been found and is given below (case (*e*)) with the corresponding values in the other cases.

Distribution: π_1 with 1 policy, π_2 with 2 policies, mean $= \mathrm{E}q$ in all cases.

Case	Variance
(*d*)	$\mathrm{N}q(\pi_1 + 4\pi_2) - \mathrm{N}q^2(\pi_1^2 + 4\pi_2^2)$
(*a*)	$\mathrm{N}q(\pi_1 + 4\pi_2) - \mathrm{N}q^2(\pi_1^2 + 4\pi_1\pi_2 + 4\pi_2^2)$
(*b*)	$\mathrm{N}q(\pi_1 + 4\pi_2) - \mathrm{N}q^2(\pi_1 + 4\pi_2)$
(*c*)	$\mathrm{N}q(\pi_1 + 4\pi_2) - \mathrm{N}q^2(\pi_1 + 4\pi_2) - 2\pi_1\pi_2 pq/(\pi_1 + 2\pi_2)^2$
(*e*)	$pq(3\mathrm{E} - 2\mathrm{N}) = \mathrm{N}q(\pi_1 + 4\pi_2) - \mathrm{N}q^2(\pi_1 + 4\pi_2)$, when $\mathrm{E} = \mathrm{N} \sum_1 t\pi_t$

It will be noted that, with the condition $\mathrm{E} = \mathrm{N} \sum_1 t\pi_t$, case (*e*) for $m = 2$ is identical with the stratified sampling case (*b*).

14. In all the above it will be noted that the probability distribution of deaths has been assumed to follow the binomial law. In our opinion the binomial or multinomial law is the appropriate law for most actuarial statistics. The Poisson law is only appropriate as an approximation to the binomial or multi-nomial law and we now propose to examine the basis of the Poisson law.

15. It arises as a limiting form of the binomial law when N tends to infinity while q tends to zero and $\mathrm{N}q$ remains finite. Implicit in the mathematical model underlying the Poisson law is the possibility of an infinity of events arising in the 'risk period' from the finite number N. To achieve this the sample must remain the same after the happening of each event or, if the event concerned means the destruction of a unit of the sample, the units destroyed must be replaced.

16. These conditions do not apply to a mortality experience if the data are in the usual form of an initial exposed to risk. After each death the population is decreased and obviously no more deaths can arise than the number exposed to risk. Sickness rates similarly do not provide a suitable application (though the

approximation is closer than for mortality rates) because a subsequent attack of sickness cannot arise while the member concerned is suffering from a previous attack. Fire insurance might be thought to provide a suitable case for the application of the Poisson law because the subject-matter of the insurance remains, provided that the claim does not totally destroy the property; but even in this case the fact that total destruction can arise shows that the theoretical conditions for Poisson variation are not fulfilled, though the practical error will be negligible. Reference may be made to the problem of counting α-particles from radioactive material; the practical conditions here are such that the Poisson law may reasonably be used though the number of items in the sample is finite.

17. There is a tendency to regard the Poisson law as more appropriate than the binomial when dealing with time rates. This idea seems to arise from the fact that the arbitrary time interval can be indefinitely reduced so that the rate for a finite time interval can be regarded as a composite rate made up of those for a large number of small time intervals. In symbols we may express the position as follows: l_0 is the number starting our finite interval; $d_0, d_1, d_2, \ldots$, etc., are the numbers of successes in the successive small intervals comprising the complete interval. These symbols are regarded as representatives of the universe so that $d_0/l_0 = q_0, d_1/l_0 = q_1, \ldots, d_t/l_0 = q_t, \ldots$ may be regarded as the probabilities corresponding to the successive small intervals. Thus, if we take a large sample N exposed to all the probabilities, each individual can be regarded as a Poisson variable and, by the reproductive property of the Poisson law, their sums will also be a Poisson variable *provided* that the variates summed are independent. It is precisely this last assumption that is not correct because, if d_0 is exceptionally large, all the other values of d are potentially affected.

18. Thus, taking mortality as an example, assuming a time interval of k small intervals and adopting the Poisson law, we can regard the survivors l_k in our universe as made up of $d_k + d_{k+1} + d_{k+2} + \ldots$ with probabilities d_k/l_0, d_{k+1}/l_0, d_{k+2}/l_0, etc. It is equally appropriate to regard each of these as Poissonian and hence, subject to independence, we can regard the sum l_k/l_0 as Poissonian. Writing $(l_0 - l_k)/l_0$ as q and l_k/l_0 as p we have the following distributions:

$$e^{-Nq}\frac{(Nq)^m}{m!} \quad \text{for the deaths}$$

and

$$e^{-Np}\frac{(Np)^n}{n!} \quad \text{for the survivors.}$$

The joint probability of m deaths and n survivors is, therefore,

$$e^{-N(p+q)}\frac{(Nq)^m(Np)^n}{m!\,n!}.$$

If we now introduce the condition that $m+n=N$, remembering that $p+q$ is necessarily equal to unity, we confine attention to those cases where

$$m = 0, 1, 2, \ldots, N \quad \text{and} \quad n = N - m.$$

The above joint probability becomes

$$e^{-N}\frac{(Nq)^m(Np)^{N-m}}{m!(N-m)!},$$

and the total value of this expression for all values of m from 0 to N is $\dfrac{e^{-N}N^N}{N!}$.

Thus the probability of m deaths becomes

$$\frac{\mathrm{N}!\,q^m p^{\mathrm{N}-m}}{m!\,(\mathrm{N}-m)!},$$

and we return full circle to the binomial distribution. The other problems can be formulated in a similar way and it is, therefore, claimed that unless the linear condition corresponding to $\Sigma m = \mathrm{N}$ does not apply the Poisson law can be adopted only as an approximation to the true binomial or multinomial law. We recognize, of course, that in suitable conditions the Poisson law is often a suitable practical assumption.

19. In view of the foregoing we cannot accept as theoretically appropriate Seal's approach to the problem of duplicates by way of a Poisson distribution. We also consider that the formulation of the problem in a form in which the number of policies on a life is a random variable serves little purpose, but there is no doubt room for difference of opinion on this aspect of the matter.

20. We also regard as extraordinary the result found by Seal in which the Central Limit Theorem does not apply even when the distribution of π_t is such that its moments are finite. This analysis has, therefore, been reworked by R. E. Beard on the basis (i) of binomial distribution of deaths and (ii) of a more critical approach to the asymptotic approximations. The following expressions represent the first four cumulants:

$$\kappa_1 = \mathrm{E}q,$$

$$\kappa_2 \sim pq\left(\mathrm{E}\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right),$$

$$\kappa_3 \sim pq(p-q)\left(\mathrm{E}\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right),$$

$$\kappa_4 \sim pq(1-6pq)\left(\mathrm{E}\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)$$
$$+3p^2q^2\left\{\left(\mathrm{E}\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)-\frac{2\sigma_2}{\sigma_1}\left(\mathrm{E}\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right)\right.$$
$$\left.+\frac{\sigma_2^2}{\sigma_1^2}\left(\mathrm{E}\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)+\left(\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)^2\right\},$$

where $\sigma_r = \Sigma t^r \pi_t$ has been used to conform with Seal's notation and $\sigma_r = m_r$ used earlier in this note.

These results are obviously related to the binomial cumulants and like them have the properties that $\kappa_3^2/\kappa_2^3 \to 0$ and $\kappa_4/\kappa_2^2 \to 0$ as $\mathrm{E} \to \infty$, thus removing the anomaly found by Seal.

21. In the Appendix the expressions for the cumulants on the basis of the Poisson law are also given; it will be found that the numerical differences between these and Seal's values are quite small.

22. Finally, consideration is given to Seal's 'fit' of the Pareto law to the data of his sample of policies on 2000 lives. He has found the parameter in the Pareto law for each age by a rough graduation of the ratios of the actual numbers of lives with 1 and 2 policies. The remaining lives with 3 or more policies represent small numbers in all cases, ranging from nil to 23. It is perhaps not

surprising, therefore, that there is an apparently good 'fit' for each age, although at some ages the long tail of his theoretical distribution, as he admits, is not representative of the data. It is, however, worth noting that the deviations (A − T) for $j = 1$ are positive in every case and, since there are twelve of them and the vast bulk of the cases are included in these groups, the 'fit' is clearly faulty.

23. It is of interest to cross-total the figures and so obtain a distribution for all ages combined. Some such grouping is essential in view of the small numbers for lives with 3 or more policies. The results and the application of the χ^2 test in this form are as follows:

j	A	T	χ^2	j	A	T	χ^2
1	1695	1649·7	1·2	9	1	3·0	
2	207	210·2	·0	10	1	2·2	
3	46	64·8	5·6	11	2	1·7	
4	22	28·7	1·7	12	—	1·4	
5	9	15·4	2·3	13	1	1·0	
6	8	9·1	·1	14	—	·9	3·0
7	4	6·0	·7	15	—	·7	
8	3	4·1	·2	16	—	·7	
				17	—	·3	
				18	1	·1	
				Total	2000	2000·0	14·8

The degrees of freedom may be taken as $9 - 2 = 7$.

24. In this form it is clear that the theoretical distribution understates the number of lives with only one policy, the balance being spread over the rest of the table with a considerable relative over-statement of the tail of the distribution. The χ^2 test does not produce such a favourable result as given by Seal. It is clear from the form of the cumulants that the final mortality distribution would be materially affected by a 'refitting' of the data to provide a distribution with a less pronounced tail. For this reason and having regard to the various points commented upon earlier we consider the numerical examples shown by Seal to be open to serious objections.

APPENDIX

BY R. E. BEARD

(A) *The problem when* E *is fixed* [see para. 10(*c*)]. In developing the expressions for the cumulants in the binomial case certain inconsistences were found in the application of the asymptotic formula. A study of the analysis suggested that the trouble arose when finding the moments about the mean from those about zero, some of the terms neglected in the asymptotic formula being of the same order as those of the required moments. Since the method found provides some insight into the nature of Darboux's approximation, the general outline is now provided.

The central problem is to determine the coefficient of u^{E} in the expansion of $\{1-\pi_1 u-\pi_2 u^2-\ldots-\pi_m u^m\}^{-k}$ when E is large. The expression can be split into partial fractions and, noting that $(1-u)$ is a factor, it takes the form

$$\begin{aligned}&\frac{\mathrm{A}_1}{(1-u)^k}+\frac{\mathrm{A}_2}{(1-u)^{k-1}}+\ldots+\frac{\mathrm{A}_k}{(1-u)}\\&+\frac{\mathrm{B}_1}{(1-\beta_1 u)^k}+\frac{\mathrm{B}_2}{(1-\beta_1 u)^{k-1}}+\ldots+\frac{\mathrm{B}_k}{(1-\beta_1 u)}\\&+\ldots\\&+\frac{\mathrm{M}_1}{(1-\beta_m u)^k}+\frac{\mathrm{M}_2}{(1-\beta_m u)^{k-1}}+\ldots+\frac{\mathrm{M}_k}{(1-\beta_m u)},\end{aligned}$$

where the A_i, B_i, ..., M_i are independent of u. Now consider the coefficient of u^{E} in the first column of terms. It is

$$\begin{aligned}&\mathrm{A}_1\frac{(\mathrm{E}+1)(\mathrm{E}+2)\ldots(\mathrm{E}+k-1)}{(k-1)!}\\&+\mathrm{B}_1\frac{(\mathrm{E}+1)(\mathrm{E}+2)\ldots(\mathrm{E}+k-1)}{(k-1)!}\beta_1^{\mathrm{E}}\\&+\ldots\\&+\mathrm{M}_1\frac{(\mathrm{E}+1)(\mathrm{E}+2)\ldots(\mathrm{E}+k-1)}{(k-1)!}\beta_m^{\mathrm{E}}.\end{aligned}$$

Now it can be shown that $\beta_1, \beta_2, \ldots, \beta_m$ (or their moduli) are all less than unity (since $\beta_r^{-1}=\alpha_r$ of Seal's paper) and hence if E is taken large enough the terms in B_1, C_1, ..., M_1 can be made as small as we please compared with that in A_1. Hence the coefficient required is the coefficient of u^{E} in

$$\frac{\mathrm{A}_1}{(1-u)^k}+\frac{\mathrm{A}_2}{(1-u)^{k-1}}+\ldots+\frac{\mathrm{A}_k}{(1-u)}$$

provided E is large. To find the $\mathrm{A}_r (r=1, 2, \ldots, k)$ consider the expression

$$\begin{aligned}\left\{\frac{1}{(1-u)\psi(u)}\right\}^k&=\left\{\frac{1}{1-\pi_1 u-\ldots-\pi_m u^m}\right\}^k\\&=\frac{\mathrm{A}_1}{(1-u)^k}+\frac{\mathrm{A}_2}{(1-u)^{k-1}}+\ldots+\frac{\mathrm{A}_k}{1-u}+\frac{\phi(u)}{\{\psi(u)\}^k},\end{aligned}$$

whence $\{\mathrm{A}_1+\mathrm{A}_2(1-u)+\ldots+\mathrm{A}_k(1-u)^{k-1}\}\{\psi(u)\}^k+(1-u)^k\phi(u)=1$. By successive differentiation with respect to u and putting $u=1$, equations for the A_r

in terms of A_1 and differential coefficients of $\{\psi(u)\}^k$ can be found. Also noting that $\psi(1)=\sigma_{(1)}$, $\psi'(1)=\frac{1}{2}\sigma_{(2)}$, $\psi''(1)=\frac{1}{3}\sigma_{(3)}$, etc., where $\psi'(1)=\{\psi'(u)\}_{u=1}$... and $\sigma_{(r)}=\Sigma t^{(r)}\pi_t$, the A_r can be found in terms of $\sigma_{(1)}$, $\sigma_{(2)}$, ... Finally, by combining these with the coefficients in the expansion of $\frac{A_1}{(1-u)^k}$, $\frac{A_2}{(1-u)^{k-1}}$, ..., the following expressions for the coefficient of u^E in $\{1-\pi_1 u-\pi_2 u^2 \ldots -\pi_m u^m\}^{-k}$ result:

k	Coeff. of u^E
1	$\frac{1}{\sigma_1}$
2	$\frac{1}{\sigma_1^2}\left(E+\frac{\sigma_2}{\sigma_1}\right)$
3	$\frac{1}{2\sigma_1^3}\left(E^2+3E\frac{\sigma_2}{\sigma_1}+\frac{3\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)$
4	$\frac{1}{6\sigma_1^4}\left[E^3+6E^2\frac{\sigma_2}{\sigma_1}+E\left\{15\left(\frac{\sigma_2}{\sigma_1}\right)^2-4\left(\frac{\sigma_3}{\sigma_1}\right)\right\}+\frac{\sigma_4}{\sigma_1}-\frac{10\sigma_2\sigma_3}{\sigma_1^2}+\frac{15\sigma_2^3}{\sigma_1^3}\right]$
5	$\frac{1}{24\sigma_1^5}\left[E^4+10E^3\frac{\sigma_2}{\sigma_1}+E^2\left\{45\left(\frac{\sigma_2}{\sigma_1}\right)^2-10\left(\frac{\sigma_3}{\sigma_1}\right)\right\}+E\left\{5\left(\frac{\sigma_4}{\sigma_1}\right)-60\frac{\sigma_2\sigma_3}{\sigma_1^2}+105\frac{\sigma_2^3}{\sigma_1^3}\right\}+105\left(\frac{\sigma_2}{\sigma_1}\right)^4+10\left(\frac{\sigma_3}{\sigma_1}\right)^2-105\frac{\sigma_2^2\sigma_3}{\sigma_1^3}+15\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right]$

The required cumulants can then be found along lines similar to those followed by Seal and finally emerge as follows:

	Binomial	Poisson
κ_1	Eq	Eq
κ_2	$pq\left(E\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)$	$q\left(E\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)$
κ_3	$pq(p-q)\left(E\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right)$	$q\left(E\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right)$
κ_4	$pq(1-6pq)\left(E\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)+3p^2q^2\left\{\left(E\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)-\frac{2\sigma_2}{\sigma_1}\left(E\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right)+\frac{\sigma_2^2}{\sigma_1^2}\left(E\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)+\left(\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)^2\right\}$	$q\left(E\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)+3q^2\left\{\left(E\frac{\sigma_4}{\sigma_1}+\frac{\sigma_2\sigma_4}{\sigma_1^2}-\frac{\sigma_5}{\sigma_1}\right)-\frac{2\sigma_2}{\sigma_1}\left(E\frac{\sigma_3}{\sigma_1}+\frac{\sigma_2\sigma_3}{\sigma_1^2}-\frac{\sigma_4}{\sigma_1}\right)+\frac{\sigma_2^2}{\sigma_1^2}\left(E\frac{\sigma_2}{\sigma_1}+\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)+\left(\frac{\sigma_2^2}{\sigma_1^2}-\frac{\sigma_3}{\sigma_1}\right)^2\right\}$

It will be noticed that these expressions are closely related to the cumulants of the binomial and Poisson distributions respectively and that $\kappa_3^2/\kappa_2^3\to 0$ and also $\kappa_4/\kappa_2^2\to 0$ as $E\to\infty$ and the peculiarity noted by Seal is due to the wrong forms

found for the cumulants. Furthermore, if $\pi_1 = 1$, $\pi_2 = \pi_3 = \ldots = 0$ these expressions reduce to the binomial and Poisson cumulants respectively.

Numerically the difference between these expressions and those found by Seal are small for the range of q and E involved, but it may be noted that

$$\sigma_3\sigma_1 > \sigma_2^2, \quad \sigma_4\sigma_1 > \sigma_2\sigma_3, \text{ etc.},$$

so that Seal's values are in excess of the true values.

(B) *On the development using a binomial instead of Poisson form.* By assuming that the probability distribution of deaths follows a binomial instead of the Poisson law a similar method of analysis can be followed, and it will be found that the moment generating function emerges as follows:

$$\frac{\text{Coeff. of } u^{\mathrm{E}} \text{ in } \left\{1 - \sum_{j=1}^{m} \pi_j (p + qe^{kj}) u^j\right\}^{-1}}{\text{Coeff. of } u^{\mathrm{E}} \text{ in } \left\{1 - \sum_{j=1}^{m} \pi_j u^j\right\}^{-1}}.$$

(C) *Distribution subject to conditions* $\Sigma n_l + \Sigma l_l = \mathrm{N}$, $\Sigma t n_l + \Sigma t l_l = \mathrm{E}$ [see paras. 12 and 13].

$$\text{M.G.F. is } \frac{\text{Coeff. of } h^{\mathrm{N}} u^{\mathrm{E}} \text{ in } \left\{1 - h \sum_{j=1}^{m} \pi_j (p + qe^{kj}) u^j\right\}^{-1}}{\text{Coeff. of } h^{\mathrm{N}} u^{\mathrm{E}} \text{ in } \left\{1 - h \sum_{j=1}^{m} \pi_j u^j\right\}^{-1}}$$

$$= \frac{\text{Coeff. of } u^{\mathrm{E}} \text{ in } \{\Sigma \pi_j (p + qe^{kj}) u^j\}^{\mathrm{N}}}{\text{Coeff. of } u^{\mathrm{E}} \text{ in } \{\Sigma \pi_j u^j\}^{\mathrm{N}}},$$

i.e., the result of dividing the Coeff. of u^{E} in

$$\{\Sigma \pi_j u^j\}^{\mathrm{N}} \left\{1 + \mathrm{N} qk \frac{\Sigma j \pi_j u^j}{\Sigma \pi_j u^j} + \frac{\mathrm{N}}{2!} qk^2 \frac{\Sigma j^2 \pi_j u^j}{\Sigma \pi_j u^j} + \frac{\mathrm{N}(\mathrm{N}-1)}{2} q^2 k^2 \left(\frac{\Sigma j \pi_j u^j}{\Sigma \pi_j u^j}\right)^2 + \ldots\right\}$$

by the Coeff. of u^{E} in $\{\Sigma \pi_j u^j\}^{\mathrm{N}}$.

When $m = 2$ it will be found that the coefficient of k is Eq and the second moment about the mean is $pq(3\mathrm{E} - 2\mathrm{N})$.

(D) *On the variance with no restraints—binomial basis* [see para. 10(*d*)]. Using Seal's notation we have

$$Pr[d_j \mid l_j; q] = \binom{l_j}{d_j} q^{d_j} (1-q)^{l_j - d_j},$$

hence
$$Pr[\{d_j\}_{j=1}^{m} \mid q] = \prod_{j=1}^{m} \binom{l_j}{d_j} q^{d_j} (1-q)^{l_j - d_j}$$

and
$$Pr\left[\sum_{j=1}^{m} j d_j = y \mid q\right] = \sum_{\Sigma j d_j = y} \prod_{j=1}^{m} \binom{l_j}{d_j} q^{d_j} (1-q)^{l_j - d_j}$$

$$= \text{Coeff. of } \alpha^y \text{ in } \Pi (p + q\alpha^j)^{l_j}.$$

Now with m groups $\pi_1, \pi_2, \ldots, \pi_m$ the joint probability of the set $\{l_j\}$ is

$$\frac{(\mathrm{N}!)^m}{\Pi(l_j)!\,\Pi(\mathrm{N} - l_j)!} \Pi (\pi_j)^{l_j} (1 - \pi_j)^{\mathrm{N} - l_j},$$

and the total probability of y deaths from the set $\{l_j\}$ is

$$Pr = \frac{(N!)^m}{\Pi(l_j)!\,\Pi(N-l_j)!}\,\Pi\left\{\frac{\pi_j(p+q\alpha^j)}{1-\pi_j}\right\}^{l_j}\Pi(1-\pi_j)^N.$$

Summing over all l_j, $l_j = 0, 1, \ldots, N$, we find

$$\begin{aligned} Pr &= \Pi\Sigma\frac{N!}{(l_j!)(N-l_j)!}\left\{\frac{\pi_j(p+q\alpha^j)}{1-\pi_j}\right\}^{l_j}(1-\pi_j)^N \\ &= \Pi\{1-\pi_j+\pi_j(p+q\alpha^j)\}^N \\ &= \Pi\{1-q\pi_j+q\pi_j\alpha^j\}^N. \end{aligned}$$

Putting $\alpha = e^k$ and expanding in powers of k we find for the M.G.F.

$$1 + Nqk\Sigma j\pi_j + \frac{k^2}{2}\{Nq\Sigma j^2\pi_j - Nq^2\Sigma j^2\pi_j^2 + N^2q^2(\Sigma j\pi_j)^2\} + \ldots,$$

whence

$$\begin{aligned} m_1 &= Nq\Sigma j\pi_j, \\ \mu_2 &= Nq\Sigma j^2\pi_j - Nq^2\Sigma j^2\pi_j^2 \\ &= Nq\Sigma j^2\pi_j - Nq^2(\Sigma j\pi_j)^2 + 2Nq^2\Sigma j\,l\pi_j\pi_l \quad (j<l). \end{aligned}$$