JIA 77 (1951) 0261-0267

DUPLICATE POLICIES IN MORTALITY DATA

By R. H. DAW, B.Sc., F.I.A.
of Powers-Samas Accounting Machines Ltd.

THE question of the effect of duplicate policies on the variance of the number of deaths in a mortality experience based on policies (not lives) has been discussed in two recent papers (Seal [1947] and Beard and Perks [1949]). The second of these papers shows that this variance depends on the sampling process envisaged and gives formulae for the variance under four such processes. It is the purpose of this paper to consider some practical aspects of the treatment of duplicates in a mortality experience.

2. Define the universe U from which the samples are drawn as consisting of an infinite number of lives (or, if preferred, a large finite number of lives and all samples made with replacement). The proportion of lives holding t policies is $\pi_t \left(t = 1, 2, \ldots, \sum_{t=1}^{\infty} \pi_t = 1\right)$. The upper limit of t must, of course, be finite, but ∞ is used to indicate that it can be large. A proportion $q(=1-p)$ of the lives in the universe are deaths, this proportion being independent of the number of policies held. The rth moment about zero of the frequency distribution π_t is denoted by m_r.

3. A sample of lives is drawn from U by some defined process and found to contain θ claims—the term 'claims' denoting the number of policies held by the lives who have died. Beard and Perks (1949) have given formulae for the variance of θ under the four sampling processes described below.

Process I. *Simple sampling of lives*

Samples of N lives are drawn at random from U:

$$\operatorname{var}\theta = Nqm_2 - Nq^2m_1^2. \tag{1}$$

Successive samples will not necessarily contain the same number of policies.

Process II. *Stratified sampling of lives*

Samples of N lives are obtained by drawing $\pi_t N$ lives at random from those lives holding t policies ($t = 1, 2, \ldots$):

$$\operatorname{var}\theta = Npqm_2. \tag{2}$$

Successive samples will contain the same number of policies, i.e. Nm_1.

Process III. *Restricted sampling of lives*

The sample is fixed at a total of E policies where $E = Nm_1$ and lives are drawn at random from U until exactly E policies are obtained:

$$\operatorname{var}\theta = Npqm_2 - pq\left(\frac{m_3}{m_1} - \frac{m_2^2}{m_1^2}\right). \tag{3}^*$$

Successive samples will not all contain N lives.

* A misprint in the denominator of the last term in formula (4) of Beard and Perks (1949) has been corrected.

Process IV. *Unrestricted sampling*

A number of random samples of N lives are drawn from U by Process I. The required sample consists of those lives with t policies from the tth sample ($t=1, 2, \ldots$):

$$\operatorname{var}\theta = Nqm_2 - Nq^2 \sum_{t=1}^{\infty} t^2\pi_t^2. \qquad (4)^*$$

Successive samples by Process IV will not all contain the same number of lives or of policies.

4. Now, consider a universe U′ which contains no duplicates (i.e. $\pi_1 = 1$), and suppose that a sample of Nm_1 lives is drawn (i.e. the sample contains as many lives as the average number of policies held by N lives in U). Then all the formulae become

$$\operatorname{var}\theta = Npqm_1, \qquad (5)$$

the usual binomial variance for a sample of Nm_1.

5. Formulae (1)–(4) all depend on the moments of the frequency distribution of the number of policies held, and before any use can be made of them this frequency distribution, or its moments, must be known. Seal (1947) examines the distribution of policies on 2000 lives and fits to this sample a discrete modification of the Pareto distribution

$$\pi_t = \frac{t^{-\beta}}{\sum_{t=1}^{\infty} t^{-\beta}} \quad (t=1, 2, \ldots). \qquad (6)$$

Because the higher moments of this distribution are infinite Seal arbitrarily terminates the series at $t=s$ policies, where s is the smallest integer satisfying

$$\frac{s^{-\beta}}{\sum_{t=1}^{\infty} t^{-\beta}} \leqslant \cdot 0004. \qquad (7)$$

Thus his formula is

$$\pi_t = \frac{t^{-\beta}}{\sum_{t=1}^{s} t^{-\beta}} \quad (t=1, 2, \ldots, s). \qquad (8)$$

6. Table 1 gives, for three such distributions, the ratio of the variances for the four sampling processes to the binomial variance of (5) for Nm_1 lives.

7. Distribution (8) has been criticized by Beard and Perks (1949) as having too long a tail to give a satisfactory description of Seal's data. Also, it by no means follows that another sample, say from the business of another office, would be distributed in any way similar to (8). Elderton (see Seal, *J.I.A.* LXXI, 41) has suggested that π_t might be a geometrical progression, i.e.

$$\pi_t = (1-r)r^{t-1} \quad (t=1, 2, \ldots), \qquad (9)$$

* Mr Perks has suggested a simpler demonstration of this formula than is given in Beard and Perks (1949). For each of the random samples of N lives drawn from U the combined probability of a death being drawn and being, also, the holder of t policies is $q\pi_t$, and therefore the variance of the number of such deaths is $Nq\pi_t(1-q\pi_t)$. Since all the samples are independent, the variance of θ is the sum of the separate variances weighted with t^2, i.e. $\sum_{t=1}^{\infty} t^2 Nq\pi_t(1-q\pi_t)$, or formula (4).

a distribution which has a less pronounced tail than the Pareto distribution. The Poisson distribution has a still shorter tail, i.e.

$$\pi_t = e^{-a}\frac{a^{t-1}}{(t-1)!} \quad (t = 1, 2, \ldots). \tag{10}$$

8. In Table 2 is shown the ratio of var θ to the binomial variance under Sampling Process II for each of the three distributions of π_t, (8), (9) and (10). For distributions (9) and (10) the parameters r and a are chosen so that the distribution has the same mean number of policies per life (or the same proportion of duplicates) as for the distributions (8) given in Table 1.

Table 1. Ratio of variance of θ to binomial variance

Sampling process	N	q	$\beta=4$	$\beta=3$	$\beta=2$
I	All	·0025	1·2675	2·0262	9·1852
	All	·01	1·2687	2·0316	9·2349
	All	·05	1·2757	2·0620	9·5131
II	All	All	1·2670	2·0244	9·1688
III	1,000	All	1·2657	2·0212	9·1310
	10,000	All	1·2669	2·0241	9·1650
	100,000	All	1·2670	2·0244	9·1684
IV	All	·0025	1·2682	2·0280	9·1912
	All	·01	1·2718	2·0390	9·2591
	All	·05	1·2922	2·1007	9·6390
Moments of distribution (8): m_1			1·1032	1·3098	2·6262
m_2			1·3978	2·6516	24·0794
Proportion of duplicates			·0936	·2365	·6192

Table 2. Ratio of variance of θ to binomial variance under Sampling Process II, and second moment of the distribution of duplicates

Mean number of policies	1·1032		1·3098		2·6262	
Proportion of duplicates	·0936		·2365		·6192	
	$\frac{\text{var }\theta}{\text{N}pqm_1}$	m_2	$\frac{\text{var }\theta}{\text{N}pqm_1}$	m_2	$\frac{\text{var }\theta}{\text{N}pqm_1}$	m_2
Distribution of duplicates:						
Discrete Pareto (8)	1·2670	1·3978	2·0244	2·6516	9·1688	24·0794
Geometrical progression (9)	1·2065	1·3311	1·6196	2·1214	4·2524	11·1679
Poisson (10)	1·1968	1·3204	1·5463	2·0254	3·2455	8·5233

9. Table 1 shows that, even if the distribution of duplicates is known exactly, it is of little practical importance which sampling process is considered the most appropriate. It also shows that the proportion of duplicates has a very considerable effect on var θ as compared with the binomial variance of (5). Further, it will be seen from Table 2 that it is not sufficient to know the

proportion of duplicates, for even when this is constant the form of the frequency distribution π_t has considerable effect on var θ.

10. At present little is known about the distribution of duplicates and their presence has often been ignored, formula (5) being used for var θ. While this may cause little practical harm when testing the graduation of a mortality table, it seems incongruous to consider applying a somewhat abstruse statistical test like the χ^2 test and to assume that formula (5) applies when at the same time it is known that var θ may be several times this value.

11. Investigations by Daw (1945) and Solomon (1948), taken in conjunction with Table 2, give some indication that the distribution of duplicates in the Assured Lives 1924–29 experience and the Continuous Mortality Investigation is more likely to be a geometrical progression or a Poisson distribution than a Pareto distribution, *provided* the 'guess' (*J.I.A.* LXVIII, 92) that the experience contains about 40% duplicates is somewhere near the truth.

12. A common graduation test is to consider

$$\frac{\theta_x - E_x q_x}{\sqrt{[E_x q_x (1 - q_x)]}} \tag{11}$$

as a unit normal deviate and to judge its significance accordingly. If the correct expression, allowing for duplicates, is

$$\frac{\theta_x - \mathrm{E}_x q_x}{\sqrt{[k\mathrm{E}_x q_x (1 - q_x)]}}, \tag{12}$$

then the use of (11) instead of (12) will result in a change in the significance level at which the test is made. Table 3 shows the significance levels actually used when (11) is applied instead of (12) at levels ·05 and ·01.

Table 3. True significance levels

k	True significance level corresponding to a significance level, when $k=1$, equal to	
	·05	·01
1·2	·07	·02
1·5	·11	·04
2·0	·17	·07
5·0	·38	·25
10·0	·54	·42

Thus even a moderate proportion of duplicates can have an appreciable effect on the significance level.

13. With all the great development which has taken place in statistical methods in the past twenty or thirty years, it is a great pity—and to some extent a criticism of the actuarial profession—that it is often not possible to apply valid statistical tests to mortality investigations without making doubtful approximations. This applies in particular to the Assured Lives 1924–29 experience and the Continuous Mortality Investigation of Assured Lives, where investigation is constantly hampered by the presence of duplicates. For example, the value of the research of Solomon (1948) is reduced since duplicates

are known to introduce heterogeneity of the type he is investigating but, because of their presence, it cannot be known whether any further element of heterogeneity is also present. Further, in applying the formulae derived by Vajda (1945) to the Continuous Mortality Investigation the presence of duplicates must be constantly borne in mind and allowed for—otherwise incorrect conclusions may be reached.

14. There are three courses which might be adopted to remedy this state of affairs:

(i) investigate the distribution of duplicates;

(ii) arrange the mortality data so that several independent estimates of each rate of mortality are obtained from which the variance can be estimated;

(iii) exclude duplicate policies and base the investigation on lives.

15. It would be laborious to investigate the distribution of duplicates for every mortality experience. If some large scale investigations were made some principle might emerge which could be applied to other mortality data, but this seems rather doubtful since the proportion will almost certainly vary with age and the form of distribution might do the same.

16. A method of obtaining independent estimates of the same rate of mortality which has been suggested by Walsh (1950) is that the exposed to risk and deaths for each age (or age-group) should be subdivided and tabulated separately according to the first letter of the surname, on the assumption that mortality is independent of this letter. There will therefore be available up to 26 sets of exposed to risk and deaths; by combining the sets into 10–15 groups with, as nearly as possible, equal numbers exposed to risk, there will be 10–15 independent estimates of the rate of mortality each of approximately equal precision. For these circumstances Walsh (1950) gives some simple significance tests which are valid whether or not duplicates are included. However, these are not the only tests which could be used; the several rates of mortality could be used to estimate the variance independently of any assumption of binomial variation and the estimate used to make valid and accurate significance tests.

17. The third course, the exclusion of duplicate policies, seems to be preferable and should not be a difficult or laborious undertaking for insurance data. Duplicates give no further information about the mortality (provided it is independent of the number of policies held) but merely have the effect of blurring the picture by increasing the random variation (i.e. the variance) of q_x.

18. If the experience from which it is desired to exclude duplicates involves other classifications than age, a number of special problems arise. It may be of interest to consider the more important of these and to take, by way of a concrete illustration, the Continuous Mortality Investigation of assured lives.

19. At present each office excludes concurrent duplicates, so there should normally be little difficulty in excluding all duplicate policies with that office, either by reference to the usual question on the proposal form about other policies held, or by any other means more suitable to the office. The exclusion of duplicate policies with other offices may be more difficult and might not be worth attempting if the effect of including them were small; some investigation of this question might be necessary.

20. In view of the subdivision of the data by class and duration of policy, consideration would have to be given to policies on the same life which belong to different classes or durations. Since the quantity of data available is large, it would probably be simplest to exclude all but the first policy on any life even if the policies did fall into different categories—otherwise any combined experience would contain duplicates.

21. However, this suggestion raises a number of difficulties. For example, it might have the effect of making too great a reduction in the quantity of data available for the select period and if so it should be possible to include duplicate policies effected in different years until the end of the select period. A difficulty arises here unless the select period can be determined before the data are collected. In the A 1924–29 experience the first 5 years of assurance were investigated separately but the final table had a select period of only 3 years. The published results of the Continuous Mortality Investigation (assured lives) also give ultimate figures for durations 3 and over and for durations 5 and over. If duplicate policies were included for 5 years, any ultimate experience excluding less than the first 5 years of assurance would contain some duplicates.

22. Consideration would also have to be given to the question whether the inclusion of one policy in the experience should lead to all subsequent policies on the same life being excluded even when the original policy has ceased to be in force; it is thought that this anomaly could be avoided by a little extra work.

23. This matter has however certain less obvious implications which will be illustrated by an example. Suppose that a life takes out an endowment assurance maturing at age 60 and then some years later, but before reaching age 60, effects a whole life policy. If the whole life policy were excluded as a duplicate and the life survived beyond age 60, then the data at the old ages of the whole life experience would be unnecessarily reduced. If it were decided that the policy with the shorter term (i.e. the endowment assurance) should be excluded then a life known to be healthy (since he has been able to effect another policy) would be excluded from the endowment assurance experience. Hence distortion of the endowment assurance experience would result since less healthy lives who proposed, but were not accepted for whole life assurances at normal terms, would be left in the experience. Similar points arise in relation to the subdivision into Medical and Non-Medical, where, it may be noted, there is a tendency for the entrants to the non-medical section to be concentrated at the younger ages to a greater extent than the medical entrants.

24. A compromise solution to some of the problems raised in the previous paragraph might be to regard, say, endowment assurances and whole life assurances as completely separate experiences when dealing with the exclusion of duplicate policies. Thus any combined experience of the various other classifications of either endowment assurances or of whole life assurances would contain no duplicates, but any amalgamation involving both endowment and whole life assurances would be liable to contain duplicates.

25. In practice, of course, considerations of administration and accuracy would have to be given due weight. How much extra work would the contributing offices be prepared to carry out and would this result in returns being delayed? Would the exclusion of duplicates render the data any less accurate or increase the liability to error?

26. Until one or other of the three courses suggested in paragraph 14 is adopted it appears to be unavoidable that statistical tests applied to much mortality data will always be subject to some unknown (and unnecessary) uncertainty in addition to the known uncertainty of the significance test. So long as that is the case, so long will statistical research into mortality be considerably hampered.

REFERENCES

BEARD, R. E. and PERKS, W. (1949). The relation between the distribution of sickness and the effect of duplicates on the distribution of deaths. *J.I.A.* LXXV, 75–86.

DAW, R. H. (1945). On the validity of statistical tests of the graduation of a mortality table. *J.I.A.* LXXII, 174–90.

SEAL, H. L. (1947). A probability distribution of deaths at age x when policies are counted instead of lives. *Skand. AktuarTidskr.* 1947, pp. 18–43.

SOLOMON, L. (1948). The analysis of heterogeneous mortality data. *J.I.A.* LXXIV, 94–112.

VAJDA, S. (1945). The analysis of variance of mortality rates. *J.I.A.* LXXII, 240–45.

WALSH, J. E. (1950). Large sample tests and confidence intervals for mortality rates. *J. Amer. Statist. Ass.* XLV, 225–37.