JIA 97 (1971) 325-330

THE APPLICATION OF THE CHI-SQUARE TEST OF GOODNESS-OF-FIT TO MORTALITY DATA GRADUATED BY SUMMATION FORMULAE

BY J. H. POLLARD, B.SC., PH.D., A.I.A.

1. INTRODUCTION

IN his paper of 1941,[3] Seal included details of some experiments he performed in an attempt to estimate the appropriate number of degrees of freedom for the chi-square goodness-of-fit test of a summation formula graduation. These results are referred to by Tetley[4] and by Benjamin and Haycocks[1] in their textbooks when they mention the difficulty of determining the number of degrees of freedom or mean chi-square value.

An approximate formula for the mean chi-square value is derived in Section 2 of this paper and extensive simulation studies in Section 4 indicate that it is very accurate. The formula is rather difficult to evaluate but fortunately a very simple approximate formula can be deduced from it, and this latter formula is accurate whenever the exposed-to-risk values $\{E_x\}$ progress in a fairly regular manner. The summation-formula method of graduation is normally only used when the exposed-to-risk values are all large and usually, under the circumstances, the $\{E_x\}$ values also progress in a fairly regular manner. Thus, the simple formula for the mean value of chi-square will be appropriate in most circumstances.

2. MATHEMATICS

Consider a population in which the unknown underlying mortality rate at age x, q_x has the shape of a third-degree polynomial. The observed mortality rate is

$$\hat{q}_x = q_x + e_x, \tag{1}$$

where e_x is the statistical error term. The exposed-to-risk at age x is denoted by E_x.

The underlying mortality rate q_x is small at most ages and it is assumed therefore that the observed number of deaths is a normal random variable with mean and variance both equal to $E_x q_x$. It follows that e_x is a normal random variable with zero mean and variance q_x/E_x.

The graduated mortality rate $\hat{q}_x$ (an estimate of q_x) is to be computed from the crude rates $\{\hat{q}_x\}$ using a summation graduation-formula with zero second-difference distortion. If the coefficients in the expanded graduation formula are denoted by $\{K_j\}$ $(j = -r, \ldots, -1, 0, 1, \ldots, r)$, the graduated mortality rate is

$$\hat{q}_x = q_x + \sum_{j=-r}^{r} K_j e_{x+j}. \tag{2}$$

The actual number of deaths observed at age x is

$$\mathrm{E}_x \mathring{q}_x = \mathrm{E}_x q_x + \mathrm{E}_x e_x \tag{3}$$

and, on the basis of the graduation, the 'expected' number of deaths is

$$\mathrm{E}_x \hat{q}_x = \mathrm{E}_x q_x + \mathrm{E}_x \left\{ \sum_{j=-r}^{r} \mathrm{K}_j e_{x+j} \right\}. \tag{4}$$

To compute the value of chi-square, we need to evaluate the sum over x of

$$\begin{aligned}
&(\mathrm{E}_x \mathring{q}_x - \mathrm{E}_x \hat{q}_x)^2 / (\mathrm{E}_x \hat{q}_x) \\
&= \frac{\mathrm{E}_x}{q_x} \left[(1-\mathrm{K}_0) e_x - \sum_{j \neq 0} \mathrm{K}_j e_{x+j} \right]^2 \left[1 + \frac{1}{q_x} \sum_{j=-r}^{r} \mathrm{K}_j e_j \right]^{-1} \\
&\doteqdot \frac{\mathrm{E}_x}{q_x} \left[(1-\mathrm{K}_0) e_x - \sum_{j \neq 0} \mathrm{K}_j e_{x+j} \right]^2 \left[1 - \frac{1}{q_x} \sum_{j=-r}^{r} \mathrm{K}_j e_j \right].
\end{aligned} \tag{5}$$

This formula will be valid provided the exposed-to-risk values are large and the q_x values are small.

The appropriate number of degrees of freedom for the chi-square test is found by examining the expected value of chi-square. The expected value of Formula (5) is

$$\begin{aligned}
&\frac{\mathrm{E}_x}{q_x} \left\{ (1-\mathrm{K}_0)^2 \mathscr{E}[e_x^2] + \sum_{j \neq 0} \mathrm{K}_j^2 \mathscr{E}[e_{x+j}^2] \right\} \\
&= \frac{\mathrm{E}_x}{q_x} \left\{ (1-\mathrm{K}_0)^2 \left[\frac{q_x}{\mathrm{E}_x} \right] + \sum_{j \neq 0} \mathrm{K}_j^2 \left[\frac{q_{x+j}}{\mathrm{E}_{x+j}} \right] \right\} \\
&= 1 - 2\mathrm{K}_0 + \sum_{j=-r}^{r} \mathrm{K}_j^2 \left(\frac{q_{x+j}}{q_x} \right) \frac{\mathrm{E}_x}{\mathrm{E}_{x+j}}.
\end{aligned} \tag{6}$$

The $\{q_x\}$ values are unknown. However, it is soon apparent that Formula (6) is fairly insensitive to the actual mortality table being investigated, and the $\{q_x\}$ values may be replaced by values $\{q_x^*\}$ from a suitable standard table or even the graduated rates $\{\hat{q}_x\}$. We conclude therefore that the appropriate number of degrees of freedom for the chi-square test on the n ages m to $m+n-1$ is

$$n(1-2\mathrm{K}_0) + \sum_{x=m}^{m+n-1} \sum_{j=-r}^{r} \mathrm{K}_j^2 \left(\frac{q_{x+j}^*}{q_x^*} \right) \frac{\mathrm{E}_x}{\mathrm{E}_{x+j}}. \tag{7}$$

Formula (7) is rather tedious to evaluate by hand and, although the calculation is straightforward on a computer, the actuary is unlikely to use it just to find the mean chi-square value. A very simple formula is available, however, and this formula is accurate whenever the exposed-to-

risk values $\{E_x\}$ progress in a fairly regular manner. It may be derived from Formula (6), noting that

$$\sum_{j=-r}^{r} K_j^2\left(\frac{q_{x+j}}{E_{x+j}}\right) \doteq \frac{q_x}{E_x} \sum_{j=-r}^{r} K_j^2 \tag{8}$$

if the exposed-to-risk values progress in a fairly regular manner. In that case, Formula (6) simplifies to

$$1-2K_0+\phi^2 \tag{9}$$

where ϕ denotes the error-reducing index of the graduation formula, and

$$\phi^2 = \sum_{j=-r}^{r} K_j^2. \tag{10}$$

We conclude that the appropriate number of degrees of freedom for a chi-square test of goodness-of-fit at n ages is

$$n[1-2K_0+\phi^2]. \tag{11}$$

This formula depends only upon the graduation formula used and the number of ages in the calculation of chi-square.

3. THE USE OF FORMULA (11)

Several assumptions were made in the derivation of Formula (11), and these merit further consideration.

The underlying q_x curve has the shape of a third-degree polynomial. This assumption will not be strictly true in practice. However, a third-degree polynomial should provide a good approximation to the q_x curve in the neighbourhood of each age x and the effects of distortion by the graduation formula will be very small.

The exposed-to-risk at each age is large. This assumption allows us to use the normal distribution rather than the binomial (or Poisson), and it also permits the negative binomial expansion in Formula (5). It is not unduly restrictive because the summation-formula method of graduation is usually used only when the exposed-to-risk is large and the crude rates progress fairly smoothly.

The underlying mortality rates $\{q_x\}$ are small. This assumption is a common one in graduation contexts and it is reasonable at all ages except the very old ones. It allows us to omit the p_x factor from the binomial variance and, when the mean chi-square value is calculated in Section 2 above, the summation of (observed–expected)^2/expected is over deaths rather than over both deaths and survivals. The exposed-to-risk at the older ages will be small and the summation graduation method is unlikely to be used at these ages.

The ratios $\{q_{x+j}/E_{x+j}\}$ progress in a fairly regular manner. This would seem

to limit the applicability of Formula (11) severely. It was noted above, however, that the summation-formula method of graduation is normally used only when the exposed-to-risk values $\{E_x\}$ are large at all ages and, usually under these circumstances, they also progress in a fairly regular manner. If the progression is far from regular, the more complicated Formula (7) should be used instead of (11).

Seal[3] performed some sampling experiments to determine the appropriate number of degrees of freedom to use with the Spencer 21-term graduation formula. He obtained graduated mortality rates at 30 consecutive ages from crude rates at 50 consecutive ages and, after nine experiments, he came to the conclusion that there were about 25 degrees of freedom. According to Formula (11), there are $30(1-\cdot34286+\cdot14320) = 24$ degrees of freedom. Details for certain other graduation formulae are given in Table 1.

It is to be noted that Formula (11) provides the correct number of degrees of freedom when a summation formula of range one is used: zero.

Table 1. *The value of the mean chi-square factor* $1-2K_0+\phi^2$ *for certain graduation formulae*

Range	Popular formula Name	Factor	Factor for optimal error-reducing formula	Factor for optimal smoothing formula
13	Spencer	·7032	·8252	·7237
15	Spencer	·7301	·8488	·7562
17	Higham	·7703	·8668	·7817
19	Larus	·7928	·8810	·8023
21	Spencer	·8003	·8925	·8193

4. A SIMULATION STUDY

Formula (11) for the mean value of chi-square is an approximate result and it would seem desirable to examine the accuracy of the formula by simulation. Hypothetical mortality experiences were built up from the fixed numbers of exposed-to-risk in Table 2 and the English Life Table No. 10 (Males) using random normal deviates which were generated by the method of Box and Muller (Hammersley and Handscomb[2]).

The fifty crude mortality rates $\{\dot{q}_x\}$ for ages 20 to 69 were then graduated by a summation formula to yield the 30 graduated rates $\{\hat{q}_x\}$ for ages 30 to 59. The *strict* chi-square value for the 10 ages 30 to 39 was calculated using the formula

$$\sum_{x=30}^{39} \frac{(E_x \dot{q}_x - E_x \hat{q}_x)^2}{E_x \hat{q}_x \hat{p}_x} \tag{12}$$

which is equivalent to

$$\sum_{x=30}^{39} \frac{(E_x \dot{q}_x - E_x \hat{q}_x)^2}{E_x \hat{q}_x} + \sum_{x=30}^{39} \frac{(E_x \dot{p}_x - E_x \hat{p}_x)^2}{E_x \hat{p}_x}. \tag{13}$$

Chi-square values were also computed for the ages 40 to 59 and 30 to 59.

The experimental mean chi-square values in Table 3 are based on 100 complete experiments and they suggest that Formula (11) is very accurate whenever the exposed-to-risk values progress in a fairly regular manner.

The effect of an irregular progression of exposed-to-risk values also needs to be studied and Table 4 contains such an irregular progression. The entries are a permutation of those in Table 2. Two hundred complete experiments were performed and the results are given in Table 5. The theoretical mean chi-square values according to Formulae (7) and (11) are also given. The theoretical values according to Formula (7) are slightly greater than those obtained from Formula (11). However, the experimental results seem to indicate that the additional labour involved in using Formula (7) is hardly worthwhile.

Table 2. *The exposed-to-risk values used in the simulation study of Formula* (11)

x	E_x	x	E_x	x	E_x	x	E_x	x	E_x
20	596	30	3,454	40	10,832	50	8,596	60	5,742
21	703	31	4,015	41	10,904	51	7,980	61	5,432
22	854	32	5,603	42	9,998	52	7,865	62	5,012
23	993	33	7,134	43	10,403	53	7,543	63	4,896
24	1,056	34	8,044	44	9,906	54	7,420	64	4,603
25	1,379	35	9,137	45	9,043	55	6,535	65	4,412
26	1,608	36	10,462	46	9,210	56	6,672	66	4,068
27	1,837	37	11,553	47	9,104	57	6,402	67	4,013
28	2,345	38	10,462	48	8,873	58	6,010	68	3,986
29	2,798	39	10,762	49	8,642	59	5,873	69	3,872

Table 3. *Mean chi-square values obtained by simulation and theoretical mean chi-square values according to Formula* (11)

	Spencer 13-term formula			Spencer 21-term formula		
	(1) 10 ages	(2) 20 ages	(3) = (1)+(2) 30 ages	(4) 10 ages	(5) 20 ages	(6) = (4)+(5) 30 ages
Theoretical Mean χ^2	7·032	14·064	21·096	8·003	16·006	24·009
Experimental Mean χ^2	7·015	13·769	20·784	8·072	15·599	23·671

5. CONCLUSION

A simple approximate formula for the mean chi-square value of a summation formula graduation is given in Section 2. This formula is shown

Table 4. *A less regular exposed-to-risk table obtained by permuting the values in Table* 2

x	E_x	x	E_x	x	E_x	x	E_x	x	E_x
20	596	30	1,379	40	3,454	50	9,137	60	10,832
21	3,872	31	4,603	41	5,873	51	7,420	61	8,642
22	703	32	1,608	42	4,015	52	10,462	62	10,904
23	3,986	33	4,896	43	6,010	53	7,543	63	8,873
24	854	34	1,837	44	5,603	54	11,553	64	9,998
25	4,013	35	5,012	45	6,402	55	7,865	65	9,104
26	993	36	2,345	46	7,134	56	10,462	66	10,403
27	4,068	37	5,432	47	6,672	57	7,980	67	9,210
28	1,056	38	2,798	48	8,044	58	10,762	68	9,906
29	4,412	39	5,742	49	6,535	59	8,596	69	9,043

Table 5. *Theoretical and experimental mean chi-square values for the irregular table of exposed-to-risk values (Table* 4)

	Spencer 21-term formula		
	(1) 10 ages	(2) 20 ages	(3) = (1)+(2) 30 ages
Theoretical Mean χ^2 (Formula (11))	8·003	16·006	24·009
Theoretical Mean χ^2 (Formula (7))	8·334	16·077	24·411
Experimental Mean χ^2 (Experiments 1–100)	8·821	17·041	25·862
Experimental Mean χ^2 (Experiments 101–200)	8·627	15·073	23·700

to be very accurate whenever the exposed-to-risk values progress in a fairly regular manner.

A more complicated formula is also given. This latter formula should be applied whenever the exposed-to-risk values progress in a rather irregular manner. Simulation studies described in Section 4 seem to indicate that the simple formula is still reasonably accurate. The additional labour involved with the more complicated formula therefore does not seem worthwhile.

REFERENCES

(1) BENJAMIN, B. AND HAYCOCKS, H. W. *The Analysis of Mortality and Other Actuarial Statistics*. Cambridge University Press, Cambridge, 1970.

(2) HAMMERSLEY, J. M. AND HANDSCOMB, D. C. *Monte Carlo Methods*. 1964.

(3) SEAL, H. L. Tests of a mortality table graduation. *J.I.A.* **71**, 5.

(4) TETLEY, H. *Actuarial Statistics*, I, 212. Cambridge University Press, Cambridge, 1964.