ESTIMATORS OF THE PROBABILITY OF THE ZERO CLASS IN POISSON AND CERTAIN RELATED POPULATIONS

By N. L. Johnson

University College, London

1. Summary and conclusions. Two estimators of the probability of falling into the zero class are compared, for a family of populations related to Poisson populations. The first estimator, ϵ_1, is based on the observed proportion in the zero class; the second, ϵ_2, would be the maximum likelihood estimator if the underlying distribution were Poisson.

From a practical point of view each estimator possesses its own peculiar advantages. ϵ_1 has the advantage that the detailed distribution among the non-zero classes need not be examined. ϵ_2 has the advantage that only the mean of the observations is needed, the distribution among the various classes not being required. The relative importance of these advantages will naturally vary according to the situations in which the estimators are to be used.

An arbitrary measure of relative accuracy, the mean square error ratio, is used. On this basis ϵ_2 is superior to ϵ_1 for all sample sizes (greater than one) if the population distribution is Poisson. Provided the sample size is not too large ϵ_2 may still be superior to ϵ_1 when the population distribution deviates to a moderate extent from Poisson form.

A third estimator ϵ_3, which is a modification of ϵ_2 and is unbiased, provided the population is Poisson, may be preferred to ϵ_2 unless p exceeds about 0.45. Its properties *vis-á-vis* ϵ_1 probably differ little from those of ϵ_2.

2. The problem. The following investigation was suggested by a problem which arose frequently in connection with the study of weapon lethality in the course of wartime operational and development research. When a fragmenting shell or bomb bursts at a given distance from a target, the density of strikes will vary according to the angular direction with regard to the equatorial plane of the shell. Within the main fragment belt, however, the density may be regarded as varying locally in a random way about an average value. The practical requirement is to determine the chance, say q, that at least one potentially lethal or effective fragment will strike an area of given size which we may call the 'unit area'. Alternatively we can estimate $p = 1 - q$, the chance that no such fragment will strike the unit area.

If it is assumed that the distribution of effective hits follows the Poisson law, and in certain cases evidence indicated that this was justifiable, then $q = 1 - e^{-m}$ and $p = e^{-m}$, where m is the expected value of the number of strikes on the unit area. It was therefore customary to estimate m from the observed average number of effective hits, $\bar{v}$ say, per unit area, derived from a series of experimental firings. Then q was estimated by the formula $1 - e^{-\bar{v}}$. If the distribution

departs from the Poisson form, the procedure is clearly incorrect in theory, but in practice the data were often inadequate to establish any alternative form of the distribution law and the estimator $1 - e^{-\bar{v}}$ was still used. In the discussion below we shall be concerned with the relative accuracy of two alternative estimators of $p(= 1 - q)$ (one of the estimators being $e^{-\bar{v}}$),

(a) when the distribution follows the Poisson law;

(b) when it departs from this law, but can be represented by a positive or negative binomial.

3. Properties of the two estimators. The problem may be stated formally as follows: $v_1, v_2, \cdots, v_n$ are independent discrete random variables. If n_0 be the number of zero values out of the n values then

$$\epsilon_1 = n_0/n \tag{1}$$

may be used as an estimator of p, the probability of the zero class. ϵ_1 is, in fact, the usual form of estimator for the proportion of individuals falling into a given class, and is of general application.

The estimator of p described in section 2 is

$$\epsilon_2 = e^{-\bar{v}}, \tag{2}$$

where $\bar{v} = n^{-1} \sum_{i=1}^{n} v_i$. This estimator is based on the assumption of a common Poisson distribution for the v's.

It will be noted that, while the evaluation of the estimator ϵ_2 does not require a knowledge of the values of the separate v's (provided their total or average is known), ϵ_1 requires only a knowledge of the number of v's which are zero. In the case described in section 2, ϵ_2 is often appropriate as the separate values of the v's are not known though their total is known. On the other hand, if, for example, $v_1, v_2, \cdots, v_n$ represent the number of cells developing in a given time in a number of cultures, it may be possible to observe only n_0, the number of cases where no development has occurred. In such cases Fisher [1] has considered the inverse problem of estimating m from n_0 by the formula $-\log \epsilon_1$. This problem will not be considered in the present paper.

We shall now compare the estimators ϵ_1 and ϵ_2 in the case when the v's do, in fact, each follow a Poisson distribution with expected value m, so that

$$Pr.\{v = r\} = \frac{m^r}{r!} e^{-m} \qquad (r = 0, 1, 2, \cdots). \tag{3}$$

The probability of the zero class is

$$p = Pr.\{v = 0\} = e^{-m}. \tag{4}$$

Since n_0 is a binomial variable with probability p and index n, the moments and moment-ratios of ϵ_1 are easily determined. In regard to ϵ_2, it can be shown that

$$\mu_s'(\epsilon_2) = p^{nf(s,n)}, \tag{5}$$

where

$$f(s, n) = 1 - e^{-s/n}. \tag{6}$$

ϵ_1 is an unbiased estimator of p while ϵ_2 is biased. Numerical calculation shows that this bias is negligible for most practical purposes (the maximum absolute bias is in the range $p = 0.3$–0.4 and is approximately $+0.18/n$). For all values of p the relation .

$$\lim_{n\to\infty} \mathfrak{E}(\epsilon_2) = p \tag{7}$$

holds.

4. Comparison of the estimators. Since ϵ_2 is a biased estimator of p, the comparison of ϵ_1 and ϵ_2 certainly cannot be based simply on their variances. One method of comparison, which does make some allowance for biases, is to use the

TABLE I

Ratio of mean square error of ϵ_2 to mean square error of ϵ_1 (Poisson population)

p \ n	10	20	30	60	∞
0.1	0.337	0.296	0.282	0.269	0.256
0.2	0.475	0.439	0.427	0.416	0.402
0.3	0.570	0.544	0.535	0.527	0.516
0.4	0.644	0.628	0.623	0.619	0.611
0.5	0.704	0.700	0.698	0.696	0.693
0.6	0.756	0.762	0.763	0.767	0.766
0.7	0.800	0.816	0.822	0.829	0.832
0.8	0.839	0.866	0.875	0.886	0.893
0.9	0.874	0.911	0.923	0.938	0.948

mean square errors of the estimators [2]. The mean square error of ϵ_2 is $\mathfrak{E}[(\epsilon_2 - p)^2] = \sigma^2(\epsilon_2) + [\mathfrak{E}(\epsilon_2) - p]^2$, while the mean square error of ϵ_1 is $\mathfrak{E}[(\epsilon_1 - p)^2] = \sigma^2(\epsilon_1)$ since ϵ_1 is an unbiased estimator of p. The ratio of mean square errors will be used as an index of comparison of estimators in the present paper, although it is clearly arbitrary, and other criteria could be preferable in certain circumstances.

Table I gives values of the mean square error ratio for various values of n and p. According to this criterion the second estimator (ϵ_2) is more accurate than the first (ϵ_1) for all cases shown in this table.

It can be shown that this ratio of mean squares must always be less than one, except in the trivial case $n = 1$. The relative advantage of ϵ_2 increases as p diminishes and does not vary greatly with n.

The correlation between the two estimators is

$$\rho(\epsilon_1, \epsilon_2) = (np)^{\frac{1}{2}}(1 - p)^{-\frac{1}{2}}\{p^{-f(1,n)} - 1\}\{p^{-n[f(1,n)]^2} - 1\}^{-\frac{1}{2}}, \tag{8}$$

whence

$$\lim_{n\to\infty} \rho(\epsilon_1, \epsilon_2) = \{-p(1-p)^{-1}\log p\}^{\frac{1}{2}}. \tag{9}$$

$\rho(\epsilon_1, \epsilon_2)$ approaches this limit rapidly as n increases. We note that

$$\lim_{n\to\infty} \rho(\epsilon_1, \epsilon_2) = \lim_{n\to\infty} (\sigma(\epsilon_2)/\sigma(\epsilon_1)), \tag{10}$$

as is to be expected since ϵ_2 is the maximum likelihood estimator of p [3].

5. A third estimator of p. The superiority of ϵ_2 as an estimator of p is to be expected, since $\bar{v}$ is a sufficient statistic for p. Using the method described in [4], we obtain the minimum variance unbiased estimator[1]

$$\epsilon_3 = (1 - n^{-1})^{n\bar{v}}, \tag{11}$$

which may be regarded as a modified, and perhaps improved, form of ϵ_2.

The variance of ϵ_3 is $p^2(p^{-1/n} - 1)$. This differs but little from the mean square error of ϵ_2, as is to be expected since $(1 - n^{-1})^n \doteq e^{-1}$. It appears that for sufficiently large values of n the mean square error of ϵ_3 will be slightly less than that of ϵ_2 for $p < 0.45$, while for $p > 0.45$ the mean square error of ϵ_2 will be slightly the smaller. The performance of ϵ_3 compared with ϵ_1 will be practically identical with that of ϵ_2.

6. Non-Poisson populations. It is quite possible that ϵ_2 (or ϵ_3) may be used as an estimator of p even when v is not in fact a Poisson variable. It may be that it has been incorrectly assumed that the distribution is Poisson in form or, perhaps, departure from Poisson, though admitted, has been considered of insufficient magnitude to affect the usefulness of ϵ_2.

It is of interest to investigate the effect of deviations from the Poisson distribution on the properties of ϵ_1 and ϵ_2. In order to do this it is first necessary to specify the nature of these deviations. Many forms of modification of the Poisson distribution have been suggested ([5]–[9]). We shall deal only with the simple form of deviation from Poisson wherein the distribution is defined by successive terms in the expansion of

$$[(1 + \omega) - \omega]^{-m/\omega}, \qquad -1 < \omega < 0 \text{ or } 0 < \omega. \tag{12}$$

The expected value of this distribution is m, whatever be the value of ω. If $-1 < \omega < 0$, then putting $\omega = -P$, $1 + \omega = Q$, $NP = m$ we have the binomial distribution

$$Pr\{v = r\} = \binom{N}{r} P^r Q^{N-r}. \tag{13}$$

[1] I am indebted to the referee for suggesting the use of this estimator.

If $0 < \omega$ we have the negative binomial distribution. Putting $\omega = 2\sigma^2$, $m = f\sigma^2$ we have

$$Pr\{v = r\} = \frac{\Gamma(r + \frac{1}{2}f)}{r!\Gamma(\frac{1}{2}f)} \cdot \frac{(2\sigma^2)^r}{(2\sigma^2 + 1)^{r+\frac{1}{2}f}}, \tag{14}$$

a form of the Pólya-Eggenberger [10] distribution previously obtained by Greenwood & Yule [11], which can be considered to arise from a mixture of Poisson distributions with expected values distributed proportionately to $\chi^2\sigma^2$ with f degrees of freedom. As $\omega \to 0$, the distribution tends to the Poisson form whether ω is moving through positive or negative values.

Whether ω is positive or negative, the probability of the zero class is

$$p = (1 + \omega)^{-m/\omega}. \tag{15}$$

The moments and moment-ratios of ϵ_1 are the same functions of p as in the Poisson case. It can be shown that

$$\mu'_s(\epsilon_2) = [1 + \omega f(s, n)]^{-mn/\omega}, \tag{16}$$

where $f(s, n) = 1 - e^{-s/n}$ as in (6), and that the correlation between the two estimators is

$$\begin{aligned}\rho(\epsilon_1, \epsilon_2) = (np)^{\frac{1}{2}}(1 - p)^{-\frac{1}{2}}&\{[1 + \omega f(1, n)]^{m/\omega} - 1\} \\ &\cdot \{[1 + \omega f(2, n)]^{-mn/\omega}[1 + \omega f(1, n)]^{2mn/\omega} - 1\}^{-\frac{1}{2}}.\end{aligned} \tag{17}$$

For any value of p, ϵ_1 is still an unbiased estimator of p, and has the same variance as when the distribution of v is Poisson. ϵ_2 is still a biased estimator of p, but the amount of bias and the variance of ϵ_2 are not the same as when the distribution of v is Poisson. Furthermore (7) no longer holds. In fact, putting $s = 1$ in (16)

$$\begin{aligned}&\mathcal{E}(\epsilon_2) = [1 + \omega(1 - e^{-1/n})]^{-mn/\omega}, \\ &\lim_{n\to\infty} \mathcal{E}(\epsilon_2) = p^{\omega/\log(1+\omega)} \neq p.\end{aligned} \tag{18}$$

7. Approximations. Since the formulae in (16) and (17) are tedious to compute, it seemed worth while investigating whether any simple approximations were possible. The following expansions in powers of n^{-1} up to the term in n^{-1} were found to give generally good results for $n \geq 30$.

$$\mathcal{E}(\epsilon_2) \doteqdot e^{-m}[1 + \tfrac{1}{2}m(1 + \omega)n^{-1}], \tag{19.1}$$

$$\sigma^2(\epsilon_2) \doteqdot e^{-2m}m(1 + \omega)n^{-1}, \tag{19.2}$$

$$\sqrt{\beta_1}(\epsilon_2) \doteqdot [nm(1 + \omega)]^{-\frac{1}{2}}[3m(1 + \omega) - (1 + 2\omega)], \tag{19.3}$$

$$\begin{aligned}\beta_2(\epsilon_2) \doteqdot 3 + 16[nm(1 + \omega)]^{-1}&[m^2(1 + \omega)^2 - 12m(1 + \omega) \\ &\cdot (1 + 2\omega) + 1 + 6\omega + 6\omega^2],\end{aligned} \tag{19.4}$$

$$\begin{aligned}\rho(\epsilon_1, \epsilon_2) \doteqdot (-\omega p \log p)^{\frac{1}{2}}&[(1 + \omega)(1 - p) \log (1 + \omega)]^{-\frac{1}{2}} \\ &\cdot [1 + (\tfrac{1}{4}m + \tfrac{1}{2}\omega - \tfrac{1}{4}m\omega)n^{-1}].\end{aligned} \tag{19.5}$$

The values of $\mathfrak{E}(\epsilon_2)$ and $\sigma^2(\epsilon_2)$ obtained from the exact formula (16) and from (19) are compared in Tables II and III respectively.

It should be noted that some of the values of ω shown do not correspond to real distributions. These cases are indicated by parentheses enclosing the corresponding figures. The values of ω chosen exhibit the trend of mathematical

TABLE II

Expected value of ϵ_2

(*Note*: The exact values and (19.1) agree to three decimal places for all cases included in this table.)

p	ω	$n = 30$	$n = 60$	$n = \infty$
0.1	−0.50	(0.193)	(0.191)	(0.190)
	−0.25	(0.139)	(0.137)	(0.135)
	0.00	0.104	0.102	0.100
	1.00	0.040	0.038	0.036
0.5	−0.25	(0.552)	(0.550)	(0.548)
	0.00	0.506	0.503	0.500
	1.00	0.380	0.374	0.368
0.9	0.00	0.902	0.901	0.900
	1.00	0.863	0.861	0.859

TABLE III

Approximate and exact values of $100\ \sigma^2(\epsilon_2)$

p	ω	$n = 30$		$n = 60$	
		Approx.	Exact	Approx.	Exact
0.1	−0.50	(0.100)	(0.104)	(0.050)	(0.051)
	−0.25	(0.091)	(0.097)	(0.046)	(0.047)
	0.00	0.077	0.083	0.038	0.040
	1.00	0.029	0.036	0.014	0.016
0.5	−0.25	(0.451)	(0.454)	(0.226)	(0.226)
	0.00	0.578	0.578	0.289	0.289
	1.00	0.901	0.910	0.451	0.451
0.9	0.00	0.284	0.277	0.142	0.140
	1.00	0.748	0.689	0.374	0.359

functions of ω which do give the moments of ϵ_2 for real distributions when ω takes certain special values, different for different p. The functions are simple continuous functions of ω and the method of presentation should not prove misleading.

Close agreement was also obtained between values given by (19.3)–(19.5) and the corresponding exact values. The approximation to $\sqrt{\beta_1}(\epsilon_2)$ was generally

correct to two decimal places and that to $\rho(\epsilon_1, \epsilon_2)$ was generally correct to three places for the values of n, ω and p in Tables II and III. $\beta_2(\epsilon_2)$ was correct to two decimal places for ω negative, while for positive ω the error did not exceed 0.04 except for $p = 0.1$ and $\omega = 1.0$ (5.09 (approx.) against 5.46).

TABLE IV

Values of $n(\omega,p)$

p \ ω	−0.2	−0.1	+0.1	+0.2
0.1	80	400	620	190
0.5	70	270	250	60
0.9	270	680	*	*

*Formula (21) gives negative values in these cases.

TABLE V

p	$f_{-2}(p)$	$f_{-1}(p)$	$f_0(p)$
0.1	5.0528	10.8992	7.5954
0.2	3.6916	6.4732	5.1006
0.3	3.1164	3.9314	3.8761
0.4	2.7809	1.6529	2.7640
0.5	2.5547	− 0.9192	1.4261
0.6	2.3889	− 4.3392	− 0.4658
0.7	2.2606	− 9.6654	− 3.5511
0.8	2.1574	−19.9286	− 9.6719
0.9	2.0722	−50.1476	−28.0060

8. A critical sample size. Using the approximate formulae (19) we see that the mean square error of ϵ_1 will be less than that of ϵ_2 provided

$$(20) \quad p(1-p)n^{-1} < (e^{-m}-p)^2 + m(1+\omega)\{e^{-m}(e^{-m}-p) + e^{-2m}\}n^{-1}.$$

This can be rewritten $n > n(\omega, p)$, where

$$(21) \quad n(\omega, p) = [p(1-p) - m(1+\omega)e^{-m}(2e^{-m}-p)](e^{-m}-p)^{-2}.$$

Provided the value of $n(\omega, p)$ given by (21) is sufficiently large for the approximation in (19) to be good, it can be said that ϵ_1 will be a better estimator of p than ϵ_2 (according to the mean square error criterion) if the sample size is bigger than $n(\omega, p)$. For smaller sample sizes it is likely that ϵ_2 will still be the superior estimator as in the Poisson case.

When $|\omega|$ is small the expansion

$$(22) \quad n(\omega, p) = \omega^{-2}f_{-2}(p) + \omega^{-1}f_{-1}(p) + f_0(p) + \cdots$$

where

$$(23.1) \quad f_{-2}(p) = 4(p \log p)^{-2}[p(1-p) + p^2 \log p],$$

$$f_{-1}(p) = 4(p \log p)^{-2}[(\tfrac{1}{3} - \tfrac{1}{2} \log p)p(1 - p) + \left(\frac{11}{6} + \log p\right)p^2 \log p], \tag{23.2}$$

$$f_0(p) = 4(p \log p)^{-2}\left[\left\{\frac{5}{48}(\log p)^2 - \frac{1}{12}\log p - \frac{1}{12}\right\} \cdot p(1 - p) + \left\{\frac{11}{48}(\log p)^2 + \frac{5}{3}\log p + \frac{5}{6}\right\} p^2 \log p\right], \tag{23.3}$$

is useful. The values of $n(\omega, p)$ given by the series (22) taken as far as $f_0(p)$ agree (to the nearest ten) with those in Table IV, which were calculated from (21). Values of $f_{-2}(p)$, $f_{-1}(p)$ and $f_0(p)$ for $p = 0.1 - (0.1) - 0.9$ are given in Table V.

REFERENCES

[1] E. JOHNSON, "Estimates of parameters by means of least squares", *Annals of Math. Stat.*, Vol. 11 (1940), p. 453.

[2] R. A. FISHER, *The Design of Experiments*, 1st ed., Oliver & Boyd, 1935, p. 221.

[3] D. BLACKWELL, "Conditional expectation and unbiased sequential estimation", *Annals of Math. Stat.*, Vol. 18 (1947), p. 105.

[4] R. A. FISHER, "Theory of statistical estimation", *Proc. Camb. Phil. Soc.*, Vol. 22 (1925), p. 700.

[5] C. V. L. CHARLIER, *Die Grundzüge der Mathematische Statistik*, Lund. Verlag Scientia, 1920, p. 80.

[6] G. PÓLYA, "Sur quelques points de la théorie des probabilités", *Ann. de l'Inst. H. Poincaré*, Vol. 1 (1931), p. 117.

[7] J. NEYMAN, "On a new class of 'contagious' distributions", *Annals of Math. Stat.*, Vol. 10 (1939), p. 35.

[8] W. FELLER, "On a general class of 'contagious' distributions", *Annals of Math. Stat.*, Vol. 14 (1943), p. 389.

[9] M. THOMAS, "A generalization of Poisson's binomial limit for use in ecology", *Biometrika*, Vol. 36 (1949), p. 18.

[10] F. EGGENBERGER AND G. PÓLYA, "Über die Statistik verketteter Vorgänge", *Zeit. für Ang. Math. und Mech.*, Vol. 1 (1923), p. 279.

[11] M. GREENWOOD AND G. YULE, "An inquiry into the nature of frequency distributions of multiple happenings", *Jour. Roy. Stat. Soc.*, Vol. 83 (1920), p. 255.