With the author's compliments

[*From* Biometrika, *Vol.* 42. *Parts* 1 *and* 2, *June* 1954.]

Substitutes for χ^2

By J. B. S. HALDANE

Department of Biometry, University College, London

Neyman (1930) and Jeffreys (1948, p. 170) have suggested a substitute for χ^2 involving some saving of computation. I here suggest what I believe to be a better one. If a sample consists of N individuals belonging to m classes, and n_r belong to the rth class, the expected number on some hypothesis being Na_r, where $\sum_{r=1}^{m} a_r = 1$, then

$$\chi^2 = \sum_{r=1}^{m} \frac{(n_r - Na_r)^2}{Na_r}.$$

Neyman's

$$\chi'^2 = \sum_{r=1}^{m} \frac{(n_r - Na_r)^2}{n_r}.$$

I consider

$$\chi''^2 = \sum_{r=1}^{m} \frac{(n_r - Na_r)^2}{n_r + 2}. \tag{1}$$

Since there is a finite probability that any n_r should be zero, it is clear that the expectation of χ'^2 is formally infinite. I shall show that it still exceeds $m-1$ even when samples in which any $n_r = 0$ are excluded. Haldane (1953) gave reasons for preferring $n_r + 1$ as a divisor in a similar context. It can be shown that

$$\mathscr{E}\left[\sum_r \frac{(n_r - Na_r)^2 + b}{n_r + c}\right] = m - 1 + N^{-1}[(b - c + 2)\sum a_r^{-1} - (3 - c)\,m + 1] + O(N^{-2}).$$

Hence to avoid an infinite expectation c must be positive, and to avoid a multiple of Σa_r^{-1}, which may be large, in the expectation, we must have $b = c - 2$. The value $b = 0$ gives a simple formula, though $b = 1$ gives an expectation nearer to $\mathscr{E}(\chi^2)$ when N is large.

Let $n_r = Na_r + x_r$. Then

$$\chi^2 = N^{-1} \sum_r x_r^2 a_r^{-1},$$

$$\chi'^2 = N^{-1} \sum_r x_r^2 a_r^{-1} \left(1 + \frac{x_r}{Na_r}\right)^{-1}$$

$$= \chi^2 + \sum_{i=2}^{\infty} [N^{-i} \sum_r (-x_r)^{i-1} a_r^{-i}],$$

$$\chi''^2 = N^{-1} \sum_r x_r^2 a_r^{-1} \left(1 + \frac{x_r + 2}{Na_r}\right)^{-1}$$

$$= \chi^2 + \sum_{i=1}^{\infty} [N^{-i-1} \sum_r x_r^2 (-x_r - 2)^i a_r^{-i-1}].$$

To find the expectations of these quantities we require the expectations of powers of x_r, namely,

$$\mathscr{E}(x_r) = 0, \quad \mathscr{E}(x_r^2) = Na_r(1-a_r), \quad \mathscr{E}(x_r^3) = Na_r(1-a_r)(1-2a_r), \quad \mathscr{E}(x_r^4) = 3N^2a_r^2(1-a_r)^2 + O(N).$$

If we write $\mathscr{E}^*(x_r^i)$ to mean the expected value of x_r^i when n_r is not zero, we omit the cases where $x_r = -Na_r$, which have a probability $(1-a_r)^N$, which tends to zero quicker than any negative power of N. Thus

$$\mathscr{E}^*(x_r) = \frac{Na_r(1-a_r)^N}{1-(1-a_r)^N}, \quad \mathscr{E}^*(x_r^2) = \frac{Na_r(1-a_r)\,1-N^2a_r^2(1-a_r)^N}{1-(1-a_r)^N}, \quad \text{etc.}$$

So

$$\left.\begin{aligned} \mathscr{E}(\chi^2) &= N^{-1}\sum_{r=1}^{m}(1-a_r) = m-1, \\ \mathscr{E}(\chi'^2) &= \infty, \\ \mathscr{E}^*(\chi'^2) &= m-1+N^{-1}(2\textstyle\sum a_r^{-1}-3m+1)+O(N^{-2}), \\ \mathscr{E}(\chi''^2) &= (m-1)\left(1-\frac{1}{N}\right)+O(N^{-2}). \end{aligned}\right\} \tag{2}$$

Thus even if we exclude the samples where any n_r is zero, χ'^2 has a positive bias often exceeding twice the reciprocal of the smallest expectation. The bias of χ''^2 is smaller, and readily calculated. The higher moments of the distribution of χ''^2 and of χ'^2, provided samples where any $n_r = 0$ are excluded, differ from those of χ^2 by quantities of the order N^{-1}. Errors of this order are neglected in the ordinary use of χ^2, and can be neglected in that of χ''^2, since χ^2 would be used if great precision were required.

As a numerical example, suppose that the numbers expected in four classes are 63, 21, 21 and 7, those observed being 71, 13, 16 and 12. Then $\chi^2 = 8{\cdot}825$, $\chi'^2 = 9{\cdot}470$, $\chi''^2 = 8{\cdot}319$. If we reverse the signs of the deviations, so that the observed numbers are 55, 29, 26 and 2, we find $\chi^2 = 8{\cdot}825$, $\chi'^2 = 16{\cdot}832$, $\chi''^2 = 10{\cdot}330$. The addition of the bias $0{\cdot}0268$ to χ''^2 gives values of $8{\cdot}345$ and $10{\cdot}357$, and this correction is clearly negligible. It is clear that χ''^2 is a far better approximation than χ'^2, and as it is no harder to calculate, it should be preferred.

REFERENCES

HALDANE, J. B. S. (1953). A class of efficient estimates of a parameter. *Bull. Int. Statist. Inst.* **33**, 231–48.

JEFFREYS, H. (1948). *Theory of Probability*. Oxford University Press.

NEYMAN, J. (1930). *Bull. Int. Statist. Inst.* **24**, 44–86.