Note on the Efficient Fitting of the Negative Binomial

R. A. Fisher

When it is desired to examine the representation of data having a_x counts of x, for values of x from 0 upward, by means of the negative binomial distribution, in which the expectation of a_x

$$E(a_x) = N\frac{(k+x-1)!}{x!(k-1)!}\frac{p^x}{(1+p)^{k+x}}$$

is expressed in terms of two parameters p and k, it is well known that the equation of estimation based on the mean

$$pk = \bar{x}$$

is fully efficient.

A second equation, with efficiency varying with the circumstances, may be taken from the second moment or variance

$$p(p+1)k = s^2$$

or, among other ways, from the frequency of zeros

$$(1+p)^k = N/a_0$$

In 1941, the author gave a number of rules (2, p. 185) for judging when the first of these is of adequate efficiency, and in 1950 (1), Anscombe has examined more fully the conditions of efficiency of both of these approaches. Many, however, will wish to use these methods only as a first step towards a fully efficient fitting, and the procedure for doing this, whatever means are used for a first orientation, is perhaps worth setting out.

Efficient scoring for k. From the primary expectation

$$m_x = E(a_x) = N\frac{(k+x-1)!}{x!(k-1)!}\cdot\frac{p^x}{(1+p)^{k+x}}$$

we have (using natural logarithms throughout)

$$\frac{\partial}{\partial p}(\log m_x) = \frac{x}{p} - \frac{k+x}{1+p}$$

whence

$$S\left\{a_x \frac{\partial}{\partial p} (\log m_x)\right\} = \frac{1}{p(1+p)} S(xa_x) - \frac{k}{1+p} S(a_x)$$
$$= \frac{N}{p(1+p)} (\bar{x} - pk)$$

If, therefore, we choose p such that

$$p = \bar{x}/k$$

the likelihood will be maximized for variation of p.

The second equation for maximum likelihood is derived from

$$\frac{\partial}{\partial k} (\log m_x) = F(k + x - 1) - F(k - 1) - \log (1 + p)$$

where $F(z)$ stands for

$$\frac{d}{dz} \log (z!)$$

and

$$F(z) - F(z - 1) = 1/z.$$

The efficient score for k is therefore

$$S\left\{a_x \frac{\partial}{\partial x} (\log m_x)\right\}$$
$$= S\left\{a_x\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{k+x-1}\right)\right\} - N \log \left(1 + \frac{\bar{x}}{k}\right)$$

In calculating the numerical value of this score for any trial value k, it is convenient first to add up the series of observations from the highest value backward, so that A_x is the number of observations exceeding x, i.e.

$$A_x = a_{x+1} + a_{x+2} + \cdots \text{ ad inf.}$$

Then the convenient expression for the score is

$$S\left(\frac{A_x}{k + x}\right) - N \log \left(1 + \frac{\bar{x}}{k}\right)$$

Trial values are then not difficult to evaluate. The value of k having maximum likelihood is $\hat{k}$, that for which the score vanishes; the corresponding value for p is $\bar{x}/\hat{k}$, and the amount of information about k is, as usual, the rate at which the score is decreasing as it passes the zero. Hence, the sampling variance and the standard deviation of the estimate may be calculated (p. 182).*

REFERENCES

(1) Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37:* 358–382, 1950.

(2) Fisher, R. A. The negative binomial distribution. *Ann. Eugenics 11:* 182–187, 1941.

* **See Bliss, C.I. (1953) Fitting the negative binomial distribution to biological data.** *Biometrics*, 9: 176-196.