Fitting the Negative Binomial Distribution to Biological Data

C. I. Bliss; R. A. Fisher

Biometrics, Vol. 9, No. 2. (Jun., 1953), pp. 176-200.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195306%299%3A2%3C176%3AFTNBDT%3E2.0.CO%3B2-L

Biometrics is currently published by International Biometric Society.

FITTING THE NEGATIVE BINOMIAL DISTRIBUTION TO BIOLOGICAL DATA

C. I. Bliss

The Connecticut Agricultural Experiment Station and Yale University

Note on the Efficient Fitting of the Negative Binomial

R. A. Fisher

Department of Genetics Cambridge

In studying the occurrence of plants and animals in nature, the number of individuals may be counted in each of many equal units of space or time. The original counts can be summarized in a frequency distribution, showing the number of units containing $x = 0, 1, 2, 3, \cdots$ individuals of a given species. If every unit in the series were exposed equally to the chance of containing the organism, the distribution would follow the Poisson series, each unit having the population mean as its expected frequency. It is easy to test whether the variation in the number of individuals per unit agrees with this hypothesis. Since the expected variance of a Poisson distribution is equal to its mean, the observed variance s^2, multiplied by the degrees of freedom n, may be divided by the sample mean $\bar{x}$ to obtain $\chi^2 = ns^2/\bar{x}$. More often than not χ^2 is significantly larger than its expectation, not only in distributions of plants and animals in nature but even in the laboratory.

A number of distributions have been devised for series in which the variance is significantly larger than the mean (2, 11, 21), frequently on the basis of more or less complex biological models. In the present paper this characteristic will be called "over dispersion". Perhaps the first of these was the negative binomial, which arose in deriving the Poisson series from the point binomial (27, 32) although it had been formulated in 1714 (2). Comparisons of expected and observed distributions have shown its wide applicability to biological data. The relative ease with which the negative binomial can be computed and

other desirable properties that have been described by Fisher (12) and by Anscombe (2) have now been supplemented by a practicable maximum likelihood estimation of its parameters described in the note by Sir Ronald Fisher at the end of the present paper. Here we will consider some characteristics of the negative binomial, estimates of its parameters and their precision, the calculation of the expected frequencies for a given sample, and its applicability to biological data.

The Negative Binomial Distribution. The negative binomial distribution is completely defined by two parameters, the arithmetic mean m and a positive exponent k. It is so called by analogy with the positive binomial distribution, $(q + p)^{n'}$, where n' is the number of individuals in a group and q and p are the expected proportions in two contrasting categories with $q + p = 1$. In computing the statistics of the positive binomial from distributions of yeast cells counted with a haemocytometer, Student (27) observed that two of his series gave negative values for p and n' but nevertheless fitted his observations very well. These and other cases (32) were described by the negative binomial, $(q - p)^{-k}$, where $p = m/k$ and $q = 1 + p$. By expansion of this expression, the probability P_x that an observational unit x will contain 0, 1, 2, $\cdots$ individuals is

$$P_x = \frac{(k + x - 1)!}{x!(k - 1)!} \cdot \frac{R^x}{q^k}, \tag{1}$$

where $R = p/q = m/(k + m)$. The probability for a given x is multiplied by N, the total number of units counted, to obtain the expected frequency (ϕ) of units with x individuals. The curve defined by the P_x's (or ϕ's) is unimodal, so that in fitting the negative binomial to an observed distribution any apparent bimodality (or multimodality) is attributed to random sampling.

The negative binomial is an extension of the Poisson series in which the population mean m, the parameter of the Poisson distribution, is not constant but varies continuously in a distribution proportional to that of χ^2. As the variance of a negative binomial approaches the mean, or the over-dispersion decreases, $k \to \infty$ and $p \to 0$. Under these conditions it can be shown (13) that the distribution converges to that for the Poisson, $P_x = e^{-m}(m^x/x!)$. Conversely, if the over-dispersion increases sufficiently, $k \to 0$. If we disregard the number of units containing no individuals, the negative binomial then converges to Fisher's logarithmic series (13), which describes effectively the apparent abundance of different species. Thus the limiting values of the exponent lead to distributions of importance in biology.

An example of the negative binomial is provided by counts of the number of European red mites on apple leaves, for which I am indebted to Dr. Garman of The Connecticut Agricultural Experiment Station (15). On July 18, 1951, 25 leaves were selected at random from each of six McIntosh trees in a single orchard receiving the same spray treatment, and the number of adult females counted on each leaf. The frequency distribution of mites on the 150 leaves is given in the first two columns of Table 1.

TABLE 1

Fitting the negative binomial to counts of red mites on apple leaves, data of P. Garman (15).

No. of mites per leaf x	No. of leaves observed f	Accumulated frequencies A_x	Expected frequencies ϕ	$\frac{(f-\phi)^2}{\phi}$
0	70	80	69.49	.004
1	38	42	37.60	.004
2	17	25	20.10	.478
3	10	15	10.70	.046
4	9	6	5.69	1.925
5	3	3	3.02	
6	2	1	1.60	.027
7	1		.85	
8+	0		.95	
Total	150 = N		150.00	2.484 = χ^2

$S(fx) = 172$, $S(fx^2) = 536$, $S(fx^3) = 2170$
$\bar{x} = 1.14667$ (Eq. 2), $s^2 = 2.27365$ (Eq. 4)

As a test of agreement with the Poisson series, the observed variance (s^2) has been computed from the basic sums beneath the table. It was nearly twice as large as the mean. From $\chi^2 = 149 \times 2.27365/1.14667 = 295.44$ with 149 degrees of freedom, P is less than .001. The over-dispersion was clearly far too large for the Poisson series. By contrast, the frequencies expected by the negative binomial, shown in the fourth column of Table 1, reproduced the observed values very closely. In the next sections, the detailed computation of the negative binomial will be illustrated with this example.

The Statistics of the Negative Binomial. The parameters m and k of the negative binomial are estimated from the frequency distribution of

a sample by the statistics $\bar{x}$ and $\hat{k}$. The mean is estimated efficiently from the frequency f of units at each x as

$$\bar{x} = S(fx)/N \tag{2}$$

For the distribution in Table 1 its numerical value was $\bar{x} = 1.146667$.

The exponent k is more difficult. Two approximations (methods 1 and 2) are available, each with a relatively high efficiency for suitable combinations of m and k (1, 2). With the maximum likelihood solution (method 3) in the note appended to this paper, either estimate may be used as a first step toward a fully efficient fitting.

(1) The original and simplest solution for the statistic k is that based upon the first and second moments (12, 32). It is determined from the mean and variance s^2 of the sample as

$$\hat{k}_1 = \frac{\bar{x}^2}{s^2 - \bar{x}}, \tag{3}$$

where, as usual, the variance is

$$s^2 = \frac{S(fx^2) - S^2(fx)/N}{N - 1} \tag{4}$$

The efficiency of the moment solution as defined by Fisher (12) has been plotted by Anscombe (1, 2) for different combinations of m and k. It has an efficiency of 90 percent or more for small values of m when $k/m > 6$, for large values of m when $k > 13$, and for m in the intermediate zone when $(k + m)(k + 2)/m \geq 15$. In practice the population values m and k are replaced necessarily by the statistics $\bar{x}$ and $\hat{k}$.

In our example, the estimation of $\hat{k}_1$ leads by Eq. 3 to $\hat{k}_1 = (1.14667)^2/(2.27365 - 1.14667) = 1.16670$. If $\hat{k}_1$ has been estimated with an efficiency of 90 percent or better, the inequality $2.313 \times 3.167/1.147 \geq 15$ should hold, but since $6.39 \not> 15$, this method is not suitable for the present series.

(2) An alternative estimate (1, 2) is based upon the ratio of the total number of units in the sample (N) to the number of units without organisms (f_0). From Eq. 1 the expected probability for $x = 0$ is $P_0 = 1/q^k$, and if P_0 is replaced by the proportion observed in the zero class, we have $f_0/N = 1/q^k$. Since $q = 1 + m/k$, an iterative solution is necessary. The required estimate of $\hat{k}_2$ is that which balances the equation

$$\hat{k}_2 \log (1 + \bar{x}/\hat{k}_2) = \log (N/f_0) \tag{5}$$

The left side of Eq. 5 is computed twice, with different trial values k', one giving a larger and the other a smaller product than the con-

stant term on the right of the equality. Interpolating between these two products for $\log(N/f_0)$ leads to a first approximation of $\hat{k}_2$, with which the process can be continued until $\hat{k}_2$ has the required precision. To estimate k with an efficiency of 90 percent or more, at least 1/3 of the units must be empty, but if the mean is less than 10, enough more empty units are required to satisfy the inequality $(m + 0.17)(P_0 - 0.32) > 0.20$. The terms in the inequality are replaced necessarily with $m = \bar{x}$ and $P_0 = f_0/N$ from the sample.

Since the number of zero frequencies in Table 1 is relatively large, method 2 is the more promising. As a test of its expected efficiency, we find $(x + 0.17)(f_0/N - 0.32) = .193$, which is very close to the level for 90 percent efficiency. From Eq. 5 the required $\hat{k}_2$ is the trial value k' for which $k' \log(1 + \bar{x}/k') = \log(150/70) = .330993$. The first trial value of $k' = 1$ was based on the estimate from method 1 and the computation arranged as follows:

k'	$1 + \bar{x}/k'$	$k' \log (1 + \bar{x}/k')$
1	2.14667	.331767
.98	2.17007	.329744
.992	2.15592	.330962

With $k_1' = 1$, the left side of Eq. 5 was too large, so that the calculation was repeated with $k_2' = .98$, reversing the inequality. By interpolation $k_3' = .98 + .02\ (.330993 - .329744)/(.331765 - .329744) = .992$, which slightly underestimated the required value. By interpolation between k_1' and k_3', $\hat{k}_2 = .99231$.

(3) For many distributions, k cannot be determined by either of the above techniques with an acceptable efficiency. In these and all critical cases, the k computed by Eq. 3 or 5 may be considered as the first step toward a definitive solution. This is provided by the method of maximum likelihood (17, 24). By suitable arrangement of the calculation, as developed by Sir Ronald Fisher in the appendix to the present paper, the procedure is practicable and rapid when the largest observation does not exceed 20 or 30. Scores (z_i) are computed from trial values of k_i', selected so that they bracket the required estimate $\hat{k}$, for which $z_i = 0$ in the equation

$$z_i = S\left(\frac{A_x}{k_i' + x}\right) - N \ln\left(1 + \frac{\bar{x}}{k_i'}\right) \tag{6}$$

where ln designates a natural logarithm. As a first step in the computation, the accumulated frequency A_x in all units containing more than

x organisms is written opposite each x. The reciprocals $1/(k_i' + x)$ from Barlow's Tables (to seven places or more) are multiplied in turn by A_x and the products accumulated to obtain the summation in the first term. The second term may be determined as the seven-place common logarithm of $(1 + \bar{x}/k_i)$ multiplied by $2.3025851 \times N$.

The first score z_1 $(i = 1)$ is computed with the first trial value k_1', usually based upon the $\hat{k}$ from Eq. 3. The second trial value k_2' depends upon the sign of z_1. If z_1 is positive, $k_2' > k_1'$; if negative, $k_2' < k_1'$, the two differing just enough to give opposite signs to z_1 and z_2. The third trial value k_3', between k_1' and k_2', is obtained by linear interpolation for $z = 0$, but the computed z_3 is rarely exactly zero. For precision, it is preferable to compute a z_4 with a sign opposite to that of z_3 by selecting k_4' at about the same distance as k_3' beyond a newly interpolated k' for $z = 0$. This provides a narrower interval within which the final $\hat{k}$ may be interpolated and a better estimate of its variance obtained.

Since the better of the two approximations of $\hat{k}$ in our example was barely 90 percent efficient, the likelihood solution would be preferred. The cumulative frequencies exceeding each x were first listed (Table 1), so that for $x = 0$, for example, $A_x = 150 - 70 = 80$, and for $x = 1$, $A_x = 80 - 38 = 42$. Starting with a trial value of $k_1' = 1.0$, the reciprocal, $1/(k_1' + x)$, for each x was multiplied by its corresponding A_x and the products accumulated in the calculator to obtain the first term in the score z (Eq. 6). This sum has been listed separately and below it the second term in the score, $345.3878 \times \log(1 + 1.146667/k')$:

	$k_1' = 1.0$	$k_2' = 1.05$	$k_3' = 1.026$	$k_4' = 1.023$
$S\{A_x/(k' + x)\}$	114.9262	110.4045	112.5247	112.7961
$-N \ln (1 + \bar{x}/k')$	−114.5875	−110.7227	−112.5432	−112.7752
z_i	.3387	−.3182	−.0185	.0209

Since z_1 was positive, the second trial k' was increased to $k_2' = 1.05$, leading to a negative z_2. Interpolating between them for $z = 0$, $k_3' = 1.0 + (.3387 \times .05)/(.3387 + .3182) = 1.026$, which, in turn, gave $z_3 = -.0185$. From linear interpolation between z_1 and z_3 for $z = 0$, $\hat{k} = 1.0247$. To insure a positive score near zero, a new trial value was selected of $k_4' = 1.023$. Interpolation between z_3 and z_4 gave the maximum likelihood estimate of $\hat{k} = 1.02459$.

The Variances of $\bar{x}$ and $\hat{k}$. The sampling variances of the statistics of the negative binomial depend upon the parameters m and k, but in practice these are replaced necessarily by the corresponding statistics

$\bar{x}$ and $\hat{k}$. The mean is computed efficiently in all cases and its variance is

$$V(\bar{x}) = \left(m + \frac{m^2}{k}\right)/N \tag{7}$$

The error variance of the mean in the example from Table 1, computed with the maximum likelihood estimate of k, was $V(\bar{x}) = (1.14667 + 1.28330)/150 = .01620$, giving the standard error $s_{\bar{x}} = .1273$.

The variance of $\hat{k}$ depends upon how it has been estimated. In general, the variance of $\hat{k}$ is less when k is small than when it is large.

(1) If computed from s^2 by Eq. 3, its large-sample variance (2) is

$$V(\hat{k}_1) \doteq \frac{2k(k+1)}{NR^2} \tag{8}$$

solved with $R = \bar{x}/(\hat{k} + \bar{x})$. For $\hat{k}_1 = 1.16670$ from Eq. 3, $R = 1.14667/2.31337 = .49567$, and $V(\hat{k}_1) = 5.0558/36.853 = .1372$, from which this estimate of k has a standard error $\sqrt{.1372} = .370$.

(2) If $\hat{k}$ is determined from the number of zero units by Eq. 5, its large-sample variance (2) is

$$V(\hat{k}_2) \doteq \frac{(1-R)^{-k} - 1 - kR}{N[-\ln(1-R) - R]^2} \tag{9}$$

where R is defined as above and ln is a natural logarithm. The error variance in the example for $\hat{k}_2 = .99231$ as estimated by Eq. 5 is solved with $R = 1.14667/2.13898 = .53608$, to obtain $V(\hat{k}_2) = .61089/8.0709 = .07569$. The standard error of $\hat{k}_2$ is 0.2751, or about 3/4 as large as that for $\hat{k}_1$.

(3) The variance of the maximum likelihood estimate of k is the reciprocal of the amount of information about k, or the rate at which the score z is decreasing as it passes the zero. It is computed, therefore, from the two values of z_i just above and below zero, say z_3 and z_4, and the two trial values of k_i' with which they have been computed, k_3' and k_4', as

$$V(\hat{k}) = \frac{k_3' - k_4'}{z_4 - z_3} \tag{10}$$

Hence the error variance of $\hat{k} = 1.02459$ is

$$V(\hat{k}) = (1.026 - 1.023)/(.0209 + .0185) = .07614,$$

so that the maximum likelihood $\hat{k}$ had a standard error of $s_{\hat{k}} = .2759$.

Tests for Agreement with the Negative Binomial. Perhaps the most convincing test of the adequacy of the negative binomial in any given case is the agreement between the frequencies (f) observed at each x and their expected values (ϕ) as computed from the statistics of the sample. The discrepancy between them is easily tested by χ^2.

The expected frequencies are computed with Eq. 1, most readily in succession and starting with the number expected at $x = 0$. This first expectation is determined with the aid of 7-place logarithms as

$$\phi_0 = N/q^k \tag{11}$$

and the succeeding entries for $x = 1, 2, 3, \cdots$ as

$$\phi_x = \frac{(k + x - 1)R}{x} \cdot \phi_{x-1} \tag{12}$$

More decimal places should be retained in the calculator at each stage than need be recorded, so as to avoid accumulating rounding errors. The observed and expected frequencies are then compare by χ^2, where

$$\chi^2 = S\left\{\frac{(f - \phi)^2}{\phi}\right\} \tag{13}$$

χ^2 has three fewer degrees of freedom than the number of ratios that are summed. As usual, the frequencies with small expectations are pooled, preferably so that no expectation is less than 5. If χ^2 shows good agreement between the matched frequencies, no other test may be needed, especially if both $\bar{x}$ and $\hat{k}$ are efficient estimates.

In the example, the expected frequency for $x = 0$ was determined with Eq. 11 as the antilog of log (150) − 1.02459 log (2.11915) or $\phi_0 = 69.4879$, giving the initial entry in the fourth column of Table 1. From $p = 1.11915$ and $q = 1 + p$, $R = 1.11915/2.11915 = .528113$ and by Eq. 12, $\phi_1 = 1.02459 \times .528113 \times 69.4879 = 37.5999$, $\phi_2 = 2.02459 \times .528113 \times 37.5999/2 = 20.1011$ and so on for successive values of x. To avoid rounding errors, six significant figures were carried in the calculation, although the ϕ's were recorded only to two decimal places. The final value, for $x = 8+$, was obtained as the difference between 150 and the sum of preceding ϕ's. The calculation of χ^2 for the discrepancy between the observed and expected frequencies (Eq. 13) is shown in the last column of Table 1, pooling the frequencies for $x \geq 5$ so as to avoid expectations of less than $\phi = 5$. The resulting $\chi^2 = 2.484$ with three degrees of freedom and $P = 0.48$ indicates good agreement with the negative binomial.

The comparison of observed and expected frequencies by χ^2 may be distorted by chance irregularities in the individual entries. Thus in

testing agreement with Neyman's contagious distribution, Beall (6) "smoothed" some of his more uneven observed frequencies before computing χ^2. This difficulty can be avoided by testing the agreement of the observed with the expected second and third moments of the negative binomial. The relation of these tests to alternative formulations, such as the logarithmic, the discrete log-normal, and the "contagious" distributions, has been considered by Anscombe (2). The moment tests have the further advantage that they take account of the few large values which are missed by grouping the tail of an observed distribution in computing χ^2.

Two tests have been described by Anscombe (2), each having the form of a difference between an observed and an "expected" moment. Although m in each test is estimated efficiently by $\bar{x}$, its variance has been derived on the assumption that an efficient estimate of k is not available for computing the expectation. A variance so derived may not apply when the expected moment is estimated by maximum likelihood. Hence the test criteria T and U are defined in terms of the estimates for which the variances are known. These variances, however, should always be computed with the best available estimate of k, usually that derived by maximum likelihood ($\hat{k}$).

The difference T between the third moment of the sample and its value predicted from the first two moments of the same sample of a negative binomial is

$$T = \frac{[x^3]}{N} - s^2\left\{\frac{2s^2}{\bar{x}} - 1\right\} \tag{13}$$

where

$$[x^3] = S\{f(x - \bar{x})^3\} = S(fx^3) - 3\bar{x}S(fx^2) + 2\bar{x}^2S(fx) \tag{14}$$

The significance of the difference T is determined by comparison with its standard error, the square root of its large-sample variance

$$V(T) = 2m(k + 1)p^2q^2[2(3 + 5p) + 3kq]/N \tag{15}$$

The variance of T should be computed with estimates of p, q and k based upon the maximum likelihood $\hat{k}$ when it is known. Although the expected third moment may be determined more accurately from the same maximum likelihood estimates as $q(q + p)m$, the variance in Eq. 15 is then of doubtful applicability.

When the observed second moment is compared with its expectation computed with $\hat{k}_2$, the difference

$$U = s^2 - (\bar{x} + \bar{x}^2/\hat{k}_2) \tag{16}$$

has the large-sample variance

$$V(U) = 2m(k+1)pq^2\left(1 - \frac{R^2}{-\ln(1-R) - R}\right) \Big/ N + p^4 V(\hat{k}_2) \quad (17)$$

$V(\hat{k}_2)$ is defined in Eq. 9 but computed with the maximum likelihood estimate $\hat{k}$ if this is known, as are the other terms in Eq. 17. Here again the expected second moment, qm, can be estimated more exactly with the maximum likelihood value for k but the applicability of the variance in Eq. 17 is then in doubt.

The differences between the observed and expected moments are much easier to compute than their standard errors. From $[x^3] = 778.47$ (Eq. 14) the present example had an observed third moment of 5.1898 and an expected value from the first two moments, $\bar{x}$ and s^2 of 6.7429, so that $T = -1.553$ by Eq. 13. Since its variance by Eq. 15 was 4.1272, the standard error of T, 2.032, showed no discrepancy from a negative binomial. The corresponding difference for the second moment and its error has been computed by Eqs. 16 and 17, to obtain $U = -0.198 \pm 0.302$, again in agreement with the negative binomial.

Models for the Negative Binomial. When the number of individuals per unit of space or time in repeated counts cannot be assumed to have the same expected value, they may represent a mixture of several homogeneous Poisson distributions. The number in each unit is restricted to the integers but this is not true of the expectations or means. In a mixture of Poissons, the means represent a positive continuous variate. The simplest frequency distribution which they might follow is the Eulerian distribution or the Pearson type III curve, and if in fact they are so distributed the observations will conform to the negative binomial.

The expected frequency may be known to vary within an observed distribution. A case in point is the distribution of bacterial clumps over a milk film (20). At least two disturbing factors were involved. In preparing a film, 0.01 milliliter of milk was placed on a microscopic slide and spread with a needle over an area of one square centimeter. Bacteria caught by surface tension on the lower surface of the drop adhered to the glass slide on contact and thus increased the concentration of bacteria in this area of the film. Secondly, the fresh drop did not have the same thickness over the entire square centimeter but due to surface tension was thicker in the center than in the margins, so that more bacteria were deposited in the center. Despite these two factors, milk meeting public health standards had so low a bacterial count that the distribution of bacterial clumps per microscopic field was seldom distinguishable from a Poisson. However, when the

TABLE 2

Observed distributions of bacterial clumps per field (Obs. f) in a milk film (20) and of yeast cells per square in a haemocytometer (27) and the expected frequencies computed from the negative binomial (Bin ϕ) and from the Neyman type A (Ney ϕ) distributions (21).

No. per unit x	Bacterial clumps Obs. f	Bacterial clumps Bin. ϕ	Yeast cells Obs. f	Yeast cells Bin. ϕ	Yeast cells Ney. ϕ
0	56	64.2	213	214.2	214.8
1	104	90.3	128	122.8	121.3
2	80	82.7	37	45.0	45.7
3	62	62.1	18	13.4	13.7
4	42	41.6	3	3.5	3.6
5	27	25.8	1	.9	.8
6	9	15.1		+.2	+.1
7	9	8.5			
8	5	4.7			
9	3	2.5			
10	2	1.3			
11		+1.2			
19	1				
N, $P(\chi^2)$	400	.54	400	.19	.18

bacterial count was high, the distribution of clumps per field reflected this known heterogeneity and then was often a negative binomial, as in the example in Table 2. This phenomenon of substantial agreement with the Poisson at low population densities and with the negative binomial at higher densities has been observed with both plant and animal populations (4, 5). A lack of randomness in microscopic counts was observed by Student in counting yeast cells with a haemocytometer (27). In fact, this seems to be the first case to be fitted with a negative binomial. The original counts are given in Table 2, together with the expected negative binomial frequencies as computed by maximum likelihood and those computed by Neyman (21) with his type A contagious distribution.

The distribution of insect pests is so seldom uniform that most experiments on insect control are randomized and replicated. In a field experiment of this type on the corn borer, four treatments were arranged in 15 randomized blocks (26). At the end of the season, eight hills of corn were selected at random in each plot and the borers recorded from each hill. This experiment has been reported both in

TABLE 3

Distribution of corn borers (Obs. f) in a field experiment arranged in 15 randomized blocks, where treatment 1 is the untreated control or check (6). The expected frequencies for the negative binomial (Bin. ϕ) have been computed independently for each treatment with the statistics in the last row of the table; those expected for the Neyman type A (Ney ϕ) are from Beall (6).

Borers per hill x	Treatment 1			Treatment 2			Treatment 3			Treatment 4		
	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ	Obs. f	Bin. ϕ	Ney. ϕ
0	19	16.6	34.4	24	19.6	22.6	43	44.3	49.8	47	45.3	53.4
1	12	18.5	6.4	16	22.2	16.7	35	31.1	23.3	23	30.1	19.7
2	18	16.9	10.4	16	19.7	18.3	17	19.1	18.9	27	18.4	17.5
3	18	14.5	11.9	18	15.9	16.4	11	11.2	12.3	9	11.0	12.1
4	11	11.9	11.2	15	12.1	13.4	5	6.4	7.3	7	6.4	7.5
5	12	9.5	9.5	9	9.0	10.3	4	3.6	4.1	3	3.7	4.4
6	7	7.5	7.9	6	6.5	7.5	1	2.0	2.2	1	2.1	2.5
7	8	5.9	6.4	5	4.6	5.2	2	1.1	1.1	1	1.2	1.4
8	4	4.5	5.2	3	3.3	3.5	2					
9	4	3.5	4.1	4	2.3	2.3		+1.2	+1.0		+1.8	+1.5
10	1	2.7	3.2	3	1.6	1.5				1		
11		2.0	2.5		1.1	.9				1		
12	1	1.5	1.9	1								
13	1	1.2	1.4		+2.1	+1.4						
15	1											
17	1	+3.3	+3.6									
19	1											
26	1											
N, $P(\chi^2)$	120	.65	.002	120	.66	.98	120	.88	.09	120	.16	.09
$\bar{x}$, k		4.033	1.532		3.167	1.764		1.483	1.333		1.508	1.190

terms of the total number of borers per plot (26) and as frequency distributions showing for each treatment the number of hills with $x = 1, 2, \cdots$ borers (6). From an analysis of variance of the plot totals (in logarithmic units) the level of borer infestation varied significantly from block to block ($P < .01$). In consequence, the composite distributions from the 15 plots for each treatment (Table 3) represented unequal levels of infestation. Negative binomials have been fitted separately to each of them. Since the expected frequencies agreed well with the observed values, the data are consistent with the hypothesis that each represented the sum of several Poisson distributions of unequal means.

The distributions in Table 3 might arise from a different model for the negative binomial in which the non-randomness is attributed to "contagion", in this case, a result of the larvae hatching from eggs that were laid in masses. Contagion, in fact, was the basis for the Neyman type A distribution developed originally for these observations (21, 6) as described in the next section. With a different mathematical formulation it leads also to a negative binomial. "Contagion" was one

TABLE 4

Distribution of the number of accidents experienced by machinists (f) and their negative binomial expectations (ϕ) (16). Distribution of soil bacteria in microscopic counts, showing the colonies per field fitted with the Poisson series, the bacteria per colony with a logarithmic series, and the bacteria per field with a negative binomial (18).

Accidents per machinist	No. of machinists f	ϕ	Colonies per field	No. of fields Obs.	Calc.	Bacteria per colony	No. of colonies Obs.	Calc.	Bacteria per field	No. of fields Obs.	Calc.
0	296	296.7	0	11	14.6	1	359	362.1	0	11	13.0
1	74	71.0	1	37	40.9	2	146	136.1	1	17	21.0
2	26	26.4	2	64	57.2	3	57	68.3	2	31	24.6
3	8	11.0	3	55	53.4	4	41	38.5	3	24	25.4
4	4	4.8	4	37	37.4	5	26	23.2	4	29	24.2
5	4	2.2	5	24	20.9	6	17	14.5	5	18	22.0
6	1	1.0	6+	12	15.6	7+	27	30.3	6	19	19.4
7		.5							7	16	16.7
8	1	.2							8	13	14.1
9		+.2							9	17	11.7
									10	6	9.6
									11	8	7.8
									12+	31	30.5
N, $P(\chi^2)$	414	.57		240	.63		673	.56		240	.52

of the explanations proposed by Student (28), who wrote, "If the presence of one individual in a division increases the chance of other individuals falling into that division, a negative binomial will fit best, but if it decreases the chance, a positive binomial". This explanation has figured prominently in the study of accident statistics (3, 16).

An early example is the distribution in Table 4 of accidents experienced by 414 machinists in three months, where the observed frequencies are matched satisfactorily by those computed with a negative binomial. If each machinist had had the same initial probability of being involved in an accident but if this probability were increased (or decreased) by his having an accident, contagion would be present and a negative binomial distribution could result. However, an opposite assumption leads to exactly the same expected distribution. If experiencing an accident had no effect upon the risk of another accident, but if the individual machinists or their shops or intervals within the three month period differed in their accident-proneness, a negative binomial would also result. Hence, the appearance of "contagion is not inherent in nature but simply in our method of sampling" (11). The relation of these two models, a mixed or compound Poisson distribution without contagion and contagion which changes the odds of further events, is discussed ably in the recent monograph by Arbous and Kerrich (3). Alternative distributions have been developed from other math-

TABLE 5

Observed distributions of quadrat counts (Obs. f) of *Lespedeza capitata* and of *Liatris aspera* in an old field association (30) and of *Primula auricula* in a grassland association (7), and their expectations with the negative binomial (Bin. ϕ), Neyman contagious type A (Ney ϕ) and Thomas double Poisson (Thom ϕ) distributions.

Plants per quadrat x	Lespedeza capitata Obs. f	Bin. ϕ	Ney* ϕ	Thom ϕ	Liatris aspera Obs. f	Bin ϕ	Thom ϕ	Primula Obs. f	Bin ϕ
0	7178	7178.1	7188.4	7279.2	7403	7403.1	7420.4	26	23.6
1	286	283.7	219.6	105.2	183	179.8	140.0	21	26.1
2	93	95.0	140.8	127.9	34	40.0	62.3	23	20.9
3	40	41.1	61.6	78.6	14	11.5	14.4	14	14.7
4	24	19.8	21.1	33.1	4	3.7	2.4	11	9.5
5	7	10.2	6.2	11.1	1	1.3	0.4	4	5.9
6	5	5.4	1.7	3.4	1	.4	0.1	5	3.5
7	1	2.9	.5	1.0		+.2		4	2.1
8	2	1.6	.1	+.5					1.2
9	1	.9						1	.7
10	2	.5							+.8
11	1	.3							
12		+.5							
N, $P(\chi^2)$	7640	.84	<.001	<.001	7640	.46	<.001	109	.70

*Computed by maximum likelihood (30), all other Neyman Type A expectations fitted by moments.

ematical definitions of "contagion" and will be considered in the next section.

The negative binomial may result from a different but related model. In counts of soil bacteria, Jones and Mollison (18) recorded for each microscopic field both the number of bacterial colonies and the number of bacteria in each colony. The number of colonies per field agreed well with the Poisson expectation for a random distribution and the number of bacteria per colony in the same counts with Fisher's logarithmic distribution. Under these conditions, the expected distribution of bacteria per field is a negative binomial (23), as confirmed by the agreement of the expected and observed frequencies (Table 4).

Quadrat counts in plant ecology which departed from Poisson have been attributed to the occurrence of plants in "clumps". Blackman (7) noted that *Primula auricula* reproduces vegetatively by short rhizomes, so that older individuals are often surrounded by younger plants. In an old-field community (30) both *Liatris astera* and *Lespedeza capitata* tended to occur in clumps. The distributions of clumps per quadrat and of plants per clump have not been reported separately, so that the model cannot be tested directly. However, the distribution of plants per quadrat in all three cases agreed excellently with the negative binomial (Table 5).

TABLE 6

Observed animal distributions (f) and their negative binomial expectations (ϕ) of *Microcalanus* nauplii in samples of marine plankton (also fitted with a Neyman type A) (5), of *Tanytarsus* in Ekman hauls (19), of Oligochaetes in Petersen hauls (19), of isopods under boards (9), and of the mite *Liponysus bacoti* on rats in Savannah (10).

Individuals per unit x	Microcalanus f	Microcalanus ϕ	Microcalanus Ney ϕ	Tanytarsus f	Tanytarsus ϕ	Oligochaetes f	Oligochaetes ϕ	Isopods f	Isopods ϕ	Mites f	Mites ϕ
0		.1	.8	32	29.5	39	34.7	28	30.2	160	160.0
1	2	.8	1.9	28	32.5	24	29.6	28	21.6	19	15.9
2	4	2.1	3.7	25	29.0	18	23.6	14	16.2	11	8.5
3	3	4.3	5.8	34	23.9	21	18.3	11	12.3	6	5.8
4	5	7.1	8.0	13	18.9	15	14.1	8	9.4	5	4.3
5	8	9.9	10.0	14	14.5	15	10.7	11	7.3	4	3.4
6	16	12.4	11.6	17	10.9	6	8.2	2	5.6	4	2.8
7	13	14.0	12.5	5	8.1	8	6.2	3	4.3	3	2.4
8	12	14.7	12.9	6	6.0	6	4.6	3	3.4	2	2.0
9	13	14.5	12.7	1	4.4	2	3.5	3	2.6	2	1.8
10	15	13.5	11.9	9	3.2	1	2.6	3	2.0		1.6
11	15	12.0	10.9		2.3	2	2.0	2	1.6		1.4
12	9	10.3	9.6		1.7	3	1.5		1.2	1	1.2
13	9	8.5	8.2	2	1.2	3	1.1	1	.9		1.1
14	7	6.8	6.8	1	.9		.8	2	.7		1.0
15	4	5.2	5.5		.6	1	.6	1	.6	2	.9
16	4	4.0	4.4		.4		+1.9		.4		.8
17	6	2.9	3.4	1	.3			2	.3		.8
18	2	2.1	2.6	1	.2				+1.4	1	.7
19		1.5	1.9		+.5					1	.6
20	2	1.1	1.4							6	+10.0
21	1	.7									
22		+1.5	+3.5								
N, $P(\chi^2)$	150	.89	.64	189	.14	164	.53	122	.34	227	.59

Other models have been developed for the negative binomial (2) and there is no reason to suppose that the possibilities have been exhausted. This is suggested in part by the variety of its applications to biological data. Some applications to fresh-water dredge samples (19) and to marine plankton (5) are shown in Table 6. It has formed the basis for a sequential sampling scheme for tapeworm cysts in whitefish (22). It has described effectively the distribution of insects in the field, including the beet leafhopper (8) and the wireworm (31), of ticks on individual sheep (12), of mites on rats (10) and of isopods under boards (9) (Table 6). Some failures in fitting can be ascribed to the inefficiency of the moment estimate of k. One of these is a count of *Ribes* on Mt. Spokane (14, 31), where Equation 3 gave $\hat{k}_1 = .134$ and Equation 6 gave $\hat{k} = .205$. Even though the estimates did not differ significantly,

TABLE 7

Distributions of quadrat counts with apparent bimodality (Obs. f) and their expected frequencies for the negative binomial (Bin ϕ), Neyman type A (Ney ϕ) and Thomas double Poisson (Thom ϕ) distributions, representing three species in a salt marsh, *Salicornia stricta*, *Plantago maritima* and *Ameria maritima* (4, 29), and a weed on arable land, *Chenopodium album* (25).

Plants per quadrat	Salicornia			Plantago			Ameria				Chenopodium		
	Obs.	Bin	Ney	Obs.	Bin	Thom	Obs.	Bin	Ney	Thom	Obs.	Bin	Thom
x	f	ϕ	ϕ	f	ϕ	ϕ	f	ϕ	ϕ	ϕ	f	ϕ	ϕ
0	4	3.3	10.7	12	7.6	11.0	57	54.1	54.9	56.4	19	9.2	19.0
1	3	6.4	4.0	8	11.3	6.7	6	16.2	7.9	5.6	5	13.5	5.0
2	8	8.4	6.5	9	12.4	10.7	12	9.0	10.1	10.0	6	14.3	9.7
3	13	9.4	7.8	13	12.0	11.2	5	5.8	9.0	9.6	9	13.0	10.6
4	11	9.6	8.1	6	10.8	10.8	5	3.9	6.5	6.8	5	11.0	9.6
5	9	9.2	7.9	8	9.3	10.0	5	2.8	4.3	4.3	20	8.8	8.3
6	8	8.4	7.6	11	7.8	8.8	7	2.0	2.8	2.8	14	6.8	7.2
7	10	7.5	7.0	7	6.4	7.4	1	1.5	1.8	1.8	8	5.2	6.0
8	3	6.5	6.4	8	5.1	6.0		1.1	1.1	1.1	4	3.8	4.9
9	3	5.6	5.7	7	4.0	4.7	1	.8	.9	.7	3	2.8	3.8
10	8	4.7	4.9	3	3.2	3.6	1	.6	.4	.4	2	2.0	3.0
11	3	3.1	4.2	4	2.5	2.7						+4.6	+7.9
12	4	2.5	3.5	1	1.9	2.0		+2.2	+.3	+.5			
13	4	2.1	2.9	1	1.4	1.4							
14		1.7	2.4										
15	3	1.3	1.9										
16		1.0	1.5	1	+4.3	+3.0							
17		.8	1.2										
18	1	.6	.9										
19				1									
20+	3	+5.9	+2.9										
N, $P(\chi^2)$	98	.48	.17	100	.14	.74	100	.014	.53	.29	95	<.001	<.001

the χ^2 test for the agreement of the observed and expected frequencies from the two estimates gave $P = .017$ and $P = .25$ respectively.

Solely as a method for summarizing a set of observations with two statistics, one of them the mean, the negative binomial should be of increasing interest to biologists. An adequate fit of this distribution to the data may serve to justify further statistical analysis such as sequential sampling (22) or a transformation for stabilizing the variance preparatory to the analysis of variance. But since several quite different models might possibly underlie data which conform to the negative binomial, one cannot use this agreement as the sole basis for justifying a particular model or conclusions based upon it.

Comparisons with Other Distributions for Over-Dispersion. Although the negative binomial is the easiest to compute and the most widely applicable of the distributions for over-dispersion, several others have been proposed. Some of these have two or more modes, while the negative binomial has only a single mode. In fitting the negative binomial to observed distributions, any "extra" modes are assumed to represent

random variation, as in fitting the negative binomial to the distributions of corn borer for treatments 1 and 2 in Table 3, of *Tanytarsus* and of *Microcalanus* in Table 6, and of quadrat counts in Table 7. In some cases, this assumption worked well, in others passably and in a few cases badly. Three two-parameter distributions have been described which may have two or more modes, the Neyman contagious type *A*, the Thomas double Poisson and the Polya. Each has been based upon a mathematical model of biological interest.

The Neyman contagious type *A* distribution (21) assumes an initial population in which groups are dispersed uniformly, within the limits of the Poisson, over the area (or period) represented by the final counts. Individuals then move out from these random centers independently but at too slow a rate to equalize their dispersion over the entire area. As numerical examples, Neyman cites the distribution of corn borers following treatment 2 (Table 3) and Student's haemocytometer counts of yeast (Table 2). In accord with theory, corn borer eggs are laid in masses and the larvae on hatching tend to migrate to neighboring corn plants. The Neyman expected frequencies, as fitted by Beall, are shown in Table 3. For treatment 2 they reproduced the observations better than the negative binomial but the reverse was true for the remaining three treatments. In the *Armeria* counts of Table 7, the Neyman type *A* again reproduced the bimodality which was missed by the negative binomial, but when fitted to unimodal distributions, the negative binomial did as well or better (Tables 2, 5, 6, 7). Fracker and Brischle (14) fitted the Neyman type *A* curve to six series of *Ribes* counts. None of them approached agreement but five of the six agreed well with the negative binomial when computed by method 1 by Wadley (31) and the sixth agreed when fitted efficiently as noted above. Despite the advantage of a potentially multimodal curve, the range of distributions fitted by the Neyman curve seems to be more restricted than with the negative binomial.

The Thomas double Poisson distribution (29) is also potentially multimodel with peaks that are somewhat more sharply defined than in the Neyman type *A*. It assumes that the number of plants per quadrat can be broken down into two Poissons, one of the number of cluster centers and the other of the number of additional plants (after the first) in a cluster. Thus individual plants of *Plantago maritima* tended to be grouped, possibly because their inflorescences are short compact spikes which frequently fall to the ground with the seeds still in the capsules (4). Frequencies computed by moments for this and other series with the Thomas distribution are given in Tables 5 and 7. They closely

resemble the expectations for the Neyman type A and have similar advantages and limitations.

Under certain conditions, the Polya distribution (2) may have two modes with somewhat larger frequencies at $x = 0$ than in the negative binomial, which in other respects it closely parallels. Under certain conditions, the Polya distribution might represent the number of individuals per quadrat in a growing population better than the negative binomial (2). If, for example, the original progeniters were released all at the same time rather than continuously over a period, the Polya distribution would be indicated, but if the individual rates of birth and death were constant and immigration occurred at a constant rate, the negative binomial would be more appropriate. So far as most biological distributions are concerned, the Polya seems to be a minor variant of the negative binomial.

A quite different alternative is provided by Fisher's logarithmic distribution (13). In studies of species, area and abundance, this has been applied effectively to the number of species (f) represented by $x = 1, 2, 3, \cdots$ individuals, in the catches of insect light traps for example (33). If k in a negative binomial approaches zero and we omit the zero class, the observed frequencies can be considered a sample from a logarithmic distribution. Williams (33) has fitted the logarithmic distribution to Buxton's data on the number of Hindu male prisoners in a south Indian jail with $x = 1, 2, 3, \cdots$ lice per head, omitting the 612 individuals (see Anscombe's correction, 2) who were free of lice. The observed frequencies agreed far better with the logarithmic expectations than with those computed from the negative binomial by method 1. In this instance, however, method 1 has an efficiency of less than 50 percent and method 2 and efficiency of nearly 98 percent. When recomputed with $\bar{x} = 6.9357 \pm .5634$ and $\hat{k}_2 = .144198 \pm .008195$, the divergence between the expected and observed frequencies (Table 8) was well within the sampling error ($\chi^2 = 31.74$, $n = 32$). Nevertheless, the observed second and third moments were significantly larger than their expectations, with $U = 243.3 \pm 38.9$ and $T = 26488 \pm 2048$. Yet, $\hat{k}_2$ was significantly larger than zero, its expected value for a logarithmic distribution.

Reexamination of the data showed that four of the 1,073 prisoners had more than 200 head lice. When $\hat{k}_2$ was recomputed without these individuals, the second and third moments were no longer discrepant, with $U = 22.86 \pm 25.68$ and $T = 575 \pm 3297$. This example indicates the disproportionate effect upon T and U of a very few large values of x. With these omissions, the expected frequencies agreed still more closely

TABLE 8

The observed distribution (f) of lice of all stages on the heads of Hindu male prisoners in Cannamore, South India, in 1937–39, and the frequencies computed with the negative binomial (Bin) from all of the data (ϕ), and from all except four prisoners having more than 200 lice (ϕ'), compared with the frequencies computed for Fisher's logarithmic distribution (Log ϕ) by omitting the 612 prisoners without lice (33, 2).

Lice per head	No. of heads				Lice per head	No. of heads			Lice per head	No. of heads		
	Obs.	Binomial		Log		Obs.	Bin	Log		Obs.	Bin	Log
x	f	ϕ	ϕ'	ϕ	x	f	ϕ	ϕ	x	f	ϕ	ϕ
0	612	612.0	612.0		11	3	9.6	8.4	22–23	8	8.3	7.0
1	106	86.5	90.5	107.2	12	10	8.7	7.6	24–25	7	7.4	6.2
2	50	48.5	50.8	52.8	13	8	8.0	6.9	26–27	10	6.6	5.6
3	29	33.9	35.5	34.7	14	6	7.4	6.3	28–29	6	6.0	5.0
4	33	26.1	27.3	25.6	15	3	6.8	5.8	30–32	2	7.9	6.7
5	20	21.2	22.1	22.2	16	6	6.3	5.4	33–35	7	6.9	5.9
6	14	17.8	18.5	16.6	17	7	5.9	5.0	36–38	9	6.0	5.2
7	12	15.3	15.8	14.0	18	4	5.5	4.6	39–41	5	5.3	4.6
8	18	13.4	13.8	12.1	19	7	5.1	4.3	42–45	5	6.1	5.3
9	11	11.9	12.2	10.6	20	7	4.8	4.1	46–49	7	5.2	4.6
10	11	10.6	10.9	9.4	21	3	4.5	3.8	50+	27	37.5	37.5

with the observed values, as evidenced by the expectations for $x = 0$ to 10 in Table 8. The agreement of the negative binomial expectations with the observed frequencies of mites on individual rats in Table 6 and of ticks on sheep as given by Fisher (12) suggest a greater usefulness for the negative binomial than for the logarithmic distribution in studies on ectoparasites.

Fitting a Single k to Several Negative Binomial Distributions. The analysis and interpretation of most biological data are facilitated by stability in the variance, so that only means need to be compared. A similar stability in the coefficient k would both increase the utility of the negative binomial and increase our confidence in its suitability for a given problem. The assumption of stability is essential in some cases as in the sequential sampling of whitefish for tapeworm cysts (22). By a simple expansion of the maximum likelihood solution, a combined $\hat{k}_c$ can be computed from a series of distributions and their homogeneity in respect to k tested by χ^2.

The calculation consists of computing the score z for each component distribution with the same trial values of k' and adding the scores for each k' over all component distributions. Trial values are selected until two sums of the scores, $S(z)$, are obtained which closely bracket 0. By interpolation between them, the required estimate $\hat{k}_c$ is that for which $S(z) = 0$. If these sums are designated as z_3 and z_4 from corresponding

trial values of k_3' and k_4', the error variance of $\hat{k}_c$ may be computed by equation 10.

The homogeneity of the k's in the component distributions depends upon the z_3 and z_4 in each individual series for k_3' and k_4'. The ratio

$$z_3^2(k_4' - k_3')/(z_3 - z_4) \tag{18}$$

is computed from each component distribution and from the totals over

TABLE 9

Calculation of a combined $\hat{k}_c$ by maximum likelihood from the four distributions of corn borer in Table 3. $N \ln(10) = 276.310212$.

Treatment No.	Term	Calculation of score with Eq. 6 for $k_1' = 1.5$	$k_2' = 1.4$	$k_3' = 1.47$	$k_4' = 1.473$	$.003\, z_3^2/(z_3 - z_4)$
1	$S\{A_x/(k' + x)\}$	156.6745	164.1457	158.8287	158.6100	
	$N \ln(1 + \bar{x}/k')$	156.6387	162.7297	158.4110	158.2317	
	z_i	.0358	1.4160	.4177	.3783	.0133
2	$S\{A_x/(k' + x)\}$	138.3464	145.2368	140.3318	140.1302	
	$N \ln(1 + \bar{x}/k')$	136.1976	141.8773	137.8480	137.6810	
	z_i	2.1488	3.3595	2.4838	2.4492	.5349
3	$S\{A_x/(k' + x)\}$	81.5615	86.2613	82.9113	82.7742	
	$N \ln(1 + \bar{x}/k')$	82.5085	86.6970	83.7206	83.5979	
	z_i	−.9470	−.4357	−.8093	−.8237	.1365
4	$S\{A_x/(k' + x)\}$	81.3606	85.9803	82.6881	82.5532	
	$N \ln(1 + \bar{x}/k')$	83.5105	87.7329	84.7322	84.6083	
	z_i	−2.1499	−1.7526	−2.0441	−2.0551	1.1395
Total	Ratios $S(z_i)$	−.9123	2.5872	.0481	−.0513	1.8242 .0001

$\chi^2 = 1.8241$

By interpolation $\hat{k}_c = 1.47145$, $V(\hat{k}_c) = .003/(.0481 + .0513) = .03018$ by Eq. 10.

all distributions. The sum of the ratios computed from the g individual distributions for k_3' and k_4' is diminished by the ratio computed from the corresponding totals of z_3 and z_4. The difference is χ^2 for testing the homogeneity of k with $g - 1$ degrees of freedom. In solving this expression, it is immaterial which boundary value, z_3 or z_4, is used, since the same χ^2 is obtained with either one.

The procedure can be illustrated with the four distributions of corn borer in Table 3, representing four different treatments in the same

field. The calculation requires for each treatment the mean $\bar{x}$ and the accumulated frequencies A_x corresponding to those in the third column of Table 1. Starting with a trial value of $k_1' = 1.5$, the calculations in Table 9 gave $S(z) = -.9123$, so that a smaller trial value, $k_2' = 1.4$, was selected next, giving $S(z) = 2.5872$. By interpolation between k_1' and k_2' for $S(z) = 0$, $k_3' = 1.47$ and from the resulting $S(z)$ by further interpolation, $k_4' = 1.473$. Interpolation between k_3' and k_4' gave the maximum likelihood estimate $\hat{k}_c$. Testing the homogeneity of k over the four treatments required the ratios in the last column of Table 9, from which $\chi^2 = 1.824$ with three degrees of freedom. We conclude that the k's for the four treatments did not differ significantly and could be represented by a single value of $\hat{k}_c = 1.4715 \pm .1737$.

Summary. In analyzing biological counts for which the variance is significantly larger than the mean, the value of the negative binomial distribution is enhanced by a simplified maximum likelihood method for estimating the coefficient k. The calculation is illustrated in detail with a numerical example and compared with other estimates of the same parameter. The error variance of the statistics of the negative binomial and tests for its agreement with a set of observations are illustrated with the same example. Models underlying the negative binomial are reviewed with reference to observed distributions of both plants and animals. A comparison with other distributions for over-dispersion suggests that the negative binomial is the most widely adaptable and generally useful of those that have been proposed so far. Finally, the maximum likelihood estimation is extended to the calculation of a single coefficient k from a series of similar distributions and the testing of their homogeneity by χ^2.

Acknowledgements. I am indebted especially to Sir Ronald Fisher, not only for the method upon which this paper depends so largely, but also for his generous guidance in applying it to many observed distributions. Dr. Philip Garman of The Connecticut Agricultural Experiment Station and Dr. E. S. Deevey, Jr. of Yale University have kindly supplied me with original data from their files, and I acknowledge with thanks the aid of Miss Theresa Santilli and Miss Margaret Robertson with the calculations.

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

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

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 (12, p. 185) for judging when the first of these is of adequate efficiency, and in 1950 (2), 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).

BIBLIOGRAPHY

(1) Anscombe, F. J. The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics 5:* 165–173, 1949.
(2) Anscombe, F. J. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37:* 358–382, 1950.
(3) Arbous, A. G. and Kerrich, J. E. Accident statistics and the concept of accident proneness. *Biometrics 7:* 340–432, 1951.
(4) Archibald, E. E. A. Plant populations. I. A new application of Neyman's contagious distribution. *Ann. Bot. 12:* 221–235, 1948.
(5) Barnes, H. and Marshall, S. M. On the variability of replicate plankton samples and some applications of "contagious" series to the statistical distribution of catches over restricted periods. *J. Marine Biol. Assoc. U. K. 30:* 233–263, 1951.
(6) Beall, G. The fit and significance of contagious distributions when applied to observations on larval insects. *Ecology 21:* 460–474, 1940.
(7) Blackman, G. E. A study by statistical methods of the distribution of species in grassland associations. *Ann. Bot. 49:* 749–777, 1935.
(8) Bowen, M. F. Population distribution of the beet leafhopper in relation to experimental field-plot lay-out. *J. Agr. Research 75:* 259–278, 1947.
(9) Cole, L. C. A theory for analyzing contagiously distributed populations. *Ecology 27:* 329–341, 1946.
(10) Cole, L. C. The measurement of interspecific association. *Ecology 30:* 411–424, 1949.
(11) Feller, W. On a general class of "contagious" distributions. *Ann. Math. Stat. 14:* 389–400, 1943.
(12) Fisher, R. A. The negative binomial distribution. *Ann. Eugenics 11:* 182–187, 1941.
(13) Fisher, R. A., Corbett, A. S. and Williams, C. B. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecology 12:* 42–58, 1943.
(14) Fracker, S. B. and Brischle, H. A. Measuring the local distribution of ribes. *Ecology 25:* 283–303, 1944.
(15) Garman, Philip. Original data on European red mite on apple leaves. Connecticut, 1951.
(16) Greenwood, M. and Yule, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Stat. Soc. 83:* 255–279, 1920.
(17) Haldane, J. B. S. The fitting of binomial distributions. *Ann. Eugenics 11:* 179–181, 1941.
(18) Jones, P. C. T., Mollison, J. E. and Quenouille, M. H. A technique for the quantitative estimation of soil microorganisms. Statistical note. *J. Gen. Microbiology 2:* 54–69, 1948.
(19) Juday, C. Unpublished data on the macroscopic fresh-water fauna in dredge samples from the bottom of Weber Lake, 1942. Courtesy of E. S. Deevey, Jr.

(20) Morgan, M. E., MacLeod, P., Anderson, E. O. and Bliss, C. I. A sequential procedure for grading milk by microscopic counts. *Storrs Agr. Expt. Sta. Bull. 276*, 1951.
(21) Neyman, J. On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Ann. Math. Stat. 10:* 35–57, 1939.
(22) Oakland, G. B. An application of sequential analysis to whitefish sampling. *Biometrics 6:* 59–67, 1950.
(23) Quenouille, M. H. A relation between the logarithmic, Poisson and negative binomial series. *Biometrics 5:* 162–164, 1949.
(24) Sichel, H. J. The estimation of the parameters of a negative binomial distribution with special reference to psychological data. *Psychometrika 16:* 107–127, 1951.
(25) Singh, B. N. and Chalam, G. V. A quantitative analysis of the weed flora on arable land. *J. Ecol. 25:* 213–221, 1937.
(26) Stirrett, G. M., Beall, G. and Timonin, M. A field experiment on the control of the European corn borer, *Pyrausta nubilalis* Hubn. by *Beauveria bassiana* Vuill. *Scient. Agric. 17:* 587–591, 1937.
(27) Student. On the error of counting with a haemocytometer. *Biometrika 5:* 351–360, 1907.
(28) Student. An explanation of deviations from Poisson's law in practice. *Biometrika 12:* 211–215, 1919.
(29) Thomas, M. A generalization of Poisson's binomial limit for use in ecology. *Biometrika 36:* 18–25, 1949.
(30) Thomson, G. W. Measures of plant aggregation based on contagious distribution. *Contr. Lab. Vert. Biol. U. of Mich. No. 53*, 1952.
(31) Wadley, F. M. Notes on the form of distribution of insect and plant populations. *Ann. Ent. Soc. Am. 43:* 581–586, 1950.
(32) Whitaker, L. On the Poisson law of small numbers. *Biometrika 10:* 36–71, 1914.
(33) Williams, C. B. Some applications of the logarithmic series and the index of diversity to ecological problems. *J. Ecology 32:* 1–44, 1944.