The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution

F. J. Anscombe

Biometrics, Vol. 5, No. 2. (Jun., 1949), pp. 165-173.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28194906%295%3A2%3C165%3ATSAOIC%3E2.0.CO%3B2-Z

Biometrics is currently published by International Biometric Society.

THE STATISTICAL ANALYSIS OF INSECT COUNTS BASED ON THE NEGATIVE BINOMIAL DISTRIBUTION

F. J. ANSCOMBE

Lecturer in Mathematics, Cambridge University, England

THIS NOTE GIVES a summary of the results of a mathematical investigation into the sampling theory of the negative binomial distribution, carried out during 1947 in the Statistical Department of Rothamsted Experimental Station. The work is a development of that of Fisher [5]. A full account will be given later elsewhere.

1. USE OF NEGATIVE BINOMIAL DISTRIBUTION

Insect counts in the field (and other population counts) are often fitted fairly well by a negative binomial distribution. This is described by two constants, the mean m and the exponent k. The variance of the distribution is

$$m + \frac{m^2}{k}, \tag{1}$$

the expected frequency of zeros is

$$p_0 = \left(1 + \frac{m}{k}\right)^{-k}, \tag{2}$$

and the chance of observing any positive count r is

$$p_r = p_0 \binom{k + r - 1}{r} \left(\frac{m}{m + k}\right)^r. \tag{3}$$

The Poisson distribution is obtained as the limit as $k \to \infty$. At the other end of the scale, as $k \to 0$, we approach the logarithmic series [6].

If we have several sets of counts on the same species of insect, from different districts or after different treatments, we may find that the mean m varies between the sets, but k remains approximately the same. To analyse such data statistically, we need to obtain a pooled estimate of k from all sets of counts and estimate the mean m separately for each set. There is some theoretical evidence [7] to show that k depends on the intrinsic power of the species to reproduce itself, while m depends on external factors. To try to fit negative binomial distributions with a common value of k to sets of counts on the same species is therefore a reasonable procedure.

2. ESTIMATION OF k FROM A SINGLE LARGE SAMPLE

We consider first the estimation of m and k from a single set of counts (made under uniform conditions). Suppose N counts have been made (i.e. the numbers of insects on N experimental units are counted), and n_0 of these counts are zeros (i.e. no insects were found on n_0 units). Let $\bar{r}$ be the average number of insects found per count (i.e. the total number of insects counted, divided by N). Then $\bar{r}$ is the best estimate of m. The best estimate of k, by the method of maximum likelihood or minimum χ^2, is tedious to find; and in practice we require a shorter method. Three methods are useful and efficient in various circumstances.

(i) We equate the variance of the sample to the variance of the distribution given above at (1). If s^2 is the sample variance, defined as the sum of squares of deviations of the N counts from $\bar{r}$, divided by $N - 1$, we get as our estimate of k

$$\frac{\bar{r}^2}{s^2 - \bar{r}} . \tag{4}$$

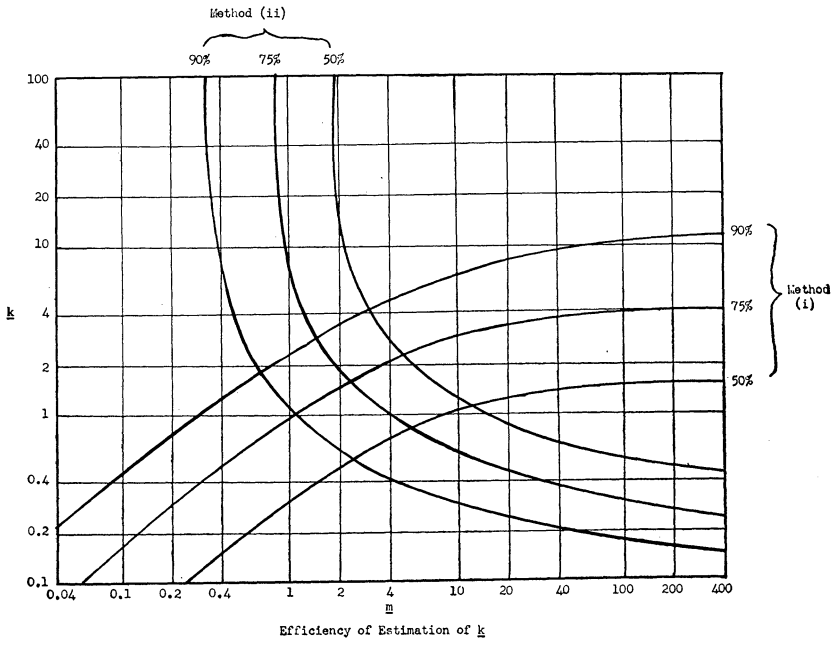
(ii) We equate the observed proportion of zeros to the expected proportion given at (2) above, i.e. we choose k by successive approximation to satisfy

$$n_0 = N\left(1 + \frac{\bar{r}}{k}\right)^{-k} . \tag{5}$$

(iii) We make a transformation of the actual counts r to a new variable y having a variance depending on k but not on m, and rather more nearly normally distributed [1]. We can then estimate k from the observed variance of y. The simplest transformation available is

$$y = \log_{10} (r + \tfrac{1}{2}k). \tag{6}$$

If k is between 2 and 5, this transformation may be used whenever m is not too small, say at least 15. If k is less than 2 or above 5, the transformation may still be used if m is large enough; but m may need to be considerably higher than 15 (so very much higher, in fact, when $k < 1$, that the possibility of use is almost ruled out). Under these conditions, the expected variance of y is approximately independent of m and equal to $0.1886\, \psi'(k)$, where $\psi'(k)$ denotes the trigamma function, i.e. the second derivative of $\ln\Gamma(k)$ with respect to



k, and has been tabulated fully by Davis [4]. Roughly, $\psi'(k) = 1/(k - \frac{1}{2})$ when k is above 2, and $1/k^2$ if k is near 0. The procedure for finding k is to guess a value, use the above transformation (6), find the sample variance s_y^2 of y, equate this to the expected variance and so get a new estimate of k. The process is repeated if the new value of k is much different from the old one.

For k not less than 2, a more elaborate transformation may be used,

$$y = \operatorname{Sinh}^{-1}\sqrt{\left(\frac{r + c}{k - 2c}\right)}. \tag{7}$$

This has an expected variance of $0.25\ \psi'(k)$. c is a constant; its best value is 0.375 if k is large, but somewhat smaller when k is small, 0.2 when $k = 2$. m may now be as low as 4 or 5. The transformation has not been investigated for $k < 2$.

Of these three methods, (i) and (ii) are quite easy to use, while (iii) is rather more bother. Roughly speaking, we use (i) if $k > 1$, (ii) if $k < 1$. The actual efficiency of the methods, as compared with the maximum

likelihood method, is indicated in the diagram, which shows, 90%, 75% and 50% efficiency contours for methods (i) and (ii). Method (iii) is only appropriate when m is not small, and then it is rather more efficient than (i).

The errors of estimation of m and k are independent, if N is large. $\bar{r}$ is always a fully efficient estimate of m.

3. ESTIMATION OF k FROM SEVERAL SAMPLES

If we have several sets of counts and wish to estimate a common value of k, we may use developments of the three methods just described.

(i) We guess a value of k, and calculate for each set of counts a quantity

$$(8) \quad T = \frac{(N-1)s^2 - (N - 1 - 1/k)\bar{r}(1 + \bar{r}/k)}{(\bar{r} + k)^2}.$$

Our object is to guess a value of k which makes the sum of these expressions T from all sets equal to zero. The process converges quite quickly if the working is suitably arranged. The divisor $(\bar{r} + k)^2$ is merely a weighting factor and can be replaced by $\bar{r}^2$ if $\bar{r}$ is always rather larger than k (this making the working easier). It is assumed here that N is not very small. Presumably 10 would be large enough, but not 2.

(ii) We guess a value of k and calculate for each set of counts a quantity

$$(9) \quad U = \log\left(1 + \frac{\bar{r}}{k}\right)\left[n_0 - \left(1 + \frac{\bar{r}}{k}\right)^{-k}\left\{N - \frac{\bar{r}(k+1)}{2(\bar{r}+k)}\right\}\right].$$

Our object is to choose a value of k which makes the sum of these expressions U for all sets equal to zero. Again, the process converges quite quickly if the working is suitably arranged. The multiplier $\log(1 + \bar{r}/k)$ is a weighting factor, and may be replaced by $\log \bar{r}$ if $\bar{r}$ is always large and much larger than k. In any case, it is not the optimum weighting factor which is much more troublesome to calculate. N is again assumed to be not very small.

(iii) We calculate the variance of the transformed variable y for each set, pool the answers and equate to the theoretical variance. The method applies even if N is as small as possible, namely 2. (No estimate of k could be

derived from a single observation.) It is, however, subject to the restrictions mentioned above on the values of m and k for a suitable transformation to exist.

4. DESIGN AND ANALYSIS OF EXPERIMENTS

Let us consider an experiment in which t treatments are compared in "randomized blocks" of tN experimental units, these units having been divided at random into t sets of N units for the application of the treatments. There may be one or several such blocks. The observations consist of counting the number of insects on each experimental unit. If negative binomial distributions with a common value of k are to be fitted to the sets of N counts for each treatment in each block, k will be estimated by one of the methods just described. The analysis will then proceed on the totals of each set of N counts. The totals have negative binomial distributions with exponent equal to Nk, and may be transformed as already explained to permit of an analysis of variance. (The transformation is different from what may have been used in determining k, since now the exponent is Nk). The residual error mean square in the analysis of variance may be compared with the expected variance for the transformation, given above, as a test of heterogeneity in the observations.

If the estimation of k is by method (iii), N may be as small as 2, but the infestations must not be too low. For methods (i) and (ii) N will need to be larger, and in the absence of more precise information it seems reasonable to recommend that N should be at least 10. These remarks amplify Beall's suggestion [3] that N should be at least 2 in experiments of this kind, so that k can be estimated.

If it is desired to avoid the assumption of a negative binomial form of distribution, with constant exponent k, it would be possible to proceed by method (iii) to derive an estimate of k and then use the transformation so defined for all further work. This would probably be as satisfactory a transformation as could be used, unless some precise assumption (other than the negative binomial one) were made about the distribution of the observations. The final analysis of variance would then be based on the totals of the individual transformed counts per set of N units and not on a direct transformation of the total count from the N units. In fact, the use of the transformation

$$y = \log (r + 1) \tag{10}$$

in this way is well known and common where the standard deviation of r appears to be roughly proportional to the mean.

TABLE

Number of eggs on ten shoots	$\bar{r}$	U (assuming $k = 0.5$)	T (assuming $k = 0.5$)
0^{10} . . . . . (27 sites)	0.0	0.00	0.0
0^9, 1 . . . . . (7 sites)	0.1	0.00	0.2
0^9, 2 . . . . . (3 sites)	0.2	0.11	3.3
0^8, 1^2	"	−0.04	−0.7
0^9, 3 . . . . . (2 sites)	0.3	0.27	7.4
0^8, 1, 4	0.5	0.36	7.5
0^8, 2, 3	"	0.36	3.5
0^6, 1^3, 2 . . . . . (3 sites)	"	−0.24	−2.5
0^5, 1^5	"	−0.54	−4.5
0^9, 6 . . . . . (2 sites)	0.6	0.87	19.1
0^6, 1^2, 2^2	"	−0.16	−2.3
0^5, 1^3, 2^2	0.7	−0.45	−3.9
0^6, 1^2, 2, 4	0.8	0.04	0.6
0^7, 2^2, 5	0.9	0.59	3.7
0^6, 1, 2, 3, 4	1.0	0.25	−0.4
0^8, 5, 6	1.1	1.36	9.5
0^7, 1, 5, 9	1.5	1.37	10.6
0^4, 1^2, 2^2, 4, 5	"	−0.43	−3.4
0^8, 3, 13	1.6	2.12	23.9
0^5, 1, 2^2, 8, 9	2.2	0.70	3.1
0^5, 2, 4, 5^2, 6	"	0.70	−3.5
0^6, 1, 3, 6, 16	2.6	1.77	12.6
0^5, 1, 2, 4^2, 17	2.8	1.11	10.9
0^9, 29	2.9	4.51	53.5
0^5, 1^2, 2, 12, 17	3.3	1.42	10.7
0^2, 1^3, 2, 4, 6, 8, 10	3.3	−1.23	−4.3
0^6, 2, 10, 11, 12	3.5	2.43	3.2
0^8, 16, 24	4.0	4.66	20.7

5. A NUMERICAL EXAMPLE

I have not encountered any experimental observations of the sort just considered. (Beall [3] gives some examples, however). To illustrate the methods of §3, I give here some counts of eggs of *Aphis fabae* made by Dr. D. Price Jones in the course of a survey of the Eastern Counties of England in 1947, which he has kindly allowed me to reproduce. Ninety-four hedgerow spindle sites were visited, that had been cut down the previous winter, so that the shoots were of one-year growth. At each site ten shoots were removed and the *A. fabae* eggs on them subsequently counted. The counts are shown in the table arranged in order of in-

TABLE (*Continued*)

Number of eggs on ten shoots	$\bar{r}$	U (assuming $k = 0.5$)	T (assuming $k = 0.5$)
0^4, 1, 2^2, 3, 7, 26	4.1	0.89	14.7
0^3, 3^2, 4, 5, 7, 9, 12	4.3	−0.01	−6.1
1, 2^2, 4^2, 5^2, 6, 8, 13	5.0	−2.92	−9.1
0^5, 1, 3, 5, 6, 40	5.5	2.49	25.2
0^2, 1^2, 2^2, 3, 8, 13, 33	6.3	−0.59	7.4
2^2, 3, 4, 5^2, 6, 9, 14^2	6.4	−2.86	−9.2
0^2, 1, 2, 3, 10, 11^2, 12, 17	6.7	−0.52	−6.5
0^2, 1, 3^2, 4, 6, 11, 15, 25	6.8	−0.51	−2.2
1, 2^3, 3, 6, 11, 14^2, 19	7.4	−2.80	−7.0
1^3, 2^3, 4, 6, 14, 47	8.0	−2.77	12.2
0^2, 1, 2, 4, 6^2, 9, 31^2	9.0	−0.17	1.0
1^3, 2, 3, 4, 10, 13, 31, 35	10.1	−2.67	−0.3
0^2, 3^2, 6, 7, 15, 19, 21, 35	10.9	0.08	−4.4
0, 3^2, 4, 5, 6, 13, 15, 25, 35	"	−1.28	−4.5
0^2, 1, 2, 4^2, 6, 23, 27, 50	11.7	0.17	3.1
0^4, 2, 3, 10, 12, 42, 50	11.9	2.98	6.8
1, 3^2, 4^2, 6, 10^2, 39, 45	12.5	−2.57	0.0
2^3, 3^2, 5, 11, 13, 18, 66	"	−2.57	7.0
0^2, 2, 6, 7, 10, 11, 12, 17, 83	14.8	0.48	9.7
0^3, 6, 10, 11, 18, 21, 35, 65	16.6	2.17	−0.9
2, 4^2, 11^2, 19, 20, 31, 32, 39	17.3	−2.41	−8.7
3, 10, 14, 15, 17^2, 23, 24^2, 33	18.0	−2.39	−11.8
0^3, 3^2, 5, 7, 13, 19, 148	19.8	2.49	31.5
4, 9, 12, 17, 18, 22, 23, 24, 34, 70	23.3	−2.25	−8.3
22, 24^2, 31, 34, 36, 43, 44, 48, 58	36.4	−2.01	−12.9
0, 1, 3, 17, 33, 38, 48, 49, 84, 110	38.3	−0.10	−5.8
1, 8, 10, 18, 26, 32, 44, 52, 82, 120	39.3	−1.97	−5.9
21, 35, 51, 59^2, 70, 105, 120, 123, 163	80.6	−1.61	−11.1

creasing $\bar{r}$. The eighth line, for example, indicates that at three sites six shoots had no eggs, three had one and one had two eggs; while the following line indicates that at one site there were five shoots without eggs and five with one egg.

The values of U and T are given on the assumption that $k = 0.5$. The sum of the Us is nearly zero, and 0.5 is close to the estimate of k given by this method. For the range of values of m that appears to have been encountered, method (ii) is considerably more efficient than method (i), while method (iii) is inappropriate. We should therefore accept the value of k given by method (ii), if any.

It appears, however, when we plot U against $\bar{r}$, that the value of k is not constant but increases with m. Thus, if we consider the counts in which $\bar{r}$ exceeds 4.0, method (ii) gives k equal to about 0.65; while for the counts in which $\bar{r}$ is less than 4.0 k is in the neighbourhood of 0.3. The effect is too marked to be attributed to the negative correlation between n_0 and $\bar{r}$ that occurs in repeated sampling of the same population. We observe a similar increase in k if we use method (i), plotting T against $\bar{r}$; but now there is further cause of perplexity, in that the values of k indicated by method (i) are appreciably lower than those of method (ii). Method (i) indicates an overall value for k round about 0.35 and 0.5 for the counts in which $\bar{r}$ exceeds 4.0. This discrepancy between methods (i) and (ii) may perhaps be due to 10 being too low a value of N for both methods to be accurate, or it may be due to a departure from the negative binomial form of distribution.

Thus, to sum up, there is clear evidence that k increases somewhat as m increases (an effect already noticed with *Myzus persicae* on potato plants [2]), and a suggestion that the form of distribution may perhaps depart from an exact negative binomial. In such an extensive series of counts, in which 940 experimental units were observed and almost 5,000 individuals (eggs) were counted, it is not surprising to find some contradiction of the simple hypothesis we started with. The same is to be expected with almost any kind of statistical material. Much attention has been given to investigating the validity of applying analysis of variance methods to yields in agricultural field experiments (without the question being entirely settled yet), and no such investigation of the validity in practice of the methods outlined in this paper has been undertaken. Our hypothesis, of negative binomial distributions with constant k, is the simplest we can make that is at all plausible; and the methods based on it are, if not elegant, at least not impossibly clumsy. It is suggested that no serious error will attend their use.

Accordingly, in further work on Price Jones's data, it would be reasonable to assume that k had a constant value of 0.5, if that facilitated the treatment. If we wished to correlate infestations at sites with other information about the sites, we could transform the total egg count per site, namely $10\bar{r}$, by the transformation

$$y = \operatorname{Sinh}^{-1}\sqrt{\left(\frac{10\bar{r} + 0.375}{4.25}\right)},$$

and treat this as a normal variable with error variance $\frac{1}{4}\psi'(5) = 0.055$. In fact, no very interesting correlations were observed, as the information about the sites was rather imprecise; and whatever associations could be perceived, visually, from scatter diagrams, were equally clear when

untransformed counts were used. However, had such counts occurred in an experiment of the sort considered in §4, much clearer correlations would be expected; and the transformation would enable treatment effects to be investigated by analysis of variance.

REFERENCES

1. Anscombe, F. J. The transformation of Poisson, Bionomial, and Negative Binomial Data. *Biometrika* 35, 246, 1948.
2. Anscombe, F. J. On Estimating the Population of Aphids in a Potato Field. *Annals of Applied Biology* 35, 567, 1948.
3. Beall, G. The Transformation of Data from Entomological Field Experiments so that the Analysis of Variance becomes Applicable. *Biometrika* 32, 243, 1942.
4. Davis, H. T. *Tables of the Higher Mathematical Functions* 2, Principia Press, Bloomington, 1935.
5. Fisher, R. A. The Negative Binomial Distribution. *Annals of Eugenics* 11, 182, 1941.
6. Fisher, R. A., Corbet, A. S., and Williams, C. B. The Relation between the Number of Individuals and the Number of Species in a Random Sample of an Animal Population. *Journal of Animal Ecology* 12, 42, 1943.
7. Kendall, D. G. On Some Modes of Population Growth Leading to R. A. Fisher's Logarithmic Series Distribution. *Biometrika* 35, 6, 1948.