193

A THEORETICAL DISTRIBUTION FOR THE APPARENT ABUNDANCE OF DIFFERENT SPECIES

Author's Note (CMS 43.53a)

The matter reprinted constitutes Part 3 (by R. A. Fisher) of a triple communication from the author, A. Steven Corbet, and C. B. Williams to the *Journal of Animal Ecology*. The tables are therefore numbered starting with Table 9.

The author would like to emphasise that the chief initiative in this discussion was taken by C. B. Williams, to whom also is due the great variety of applications in which the distribution and the discussion have been found useful. The author has been concerned only with establishing the relationship of the new distribution to others previously studied, notably the Poisson series and the negative binomial; to demonstrating the fundamental mathematical relationships; to providing tables of sufficient accuracy and range to facilitate to the utmost the numerical calculations which workers were inclined to make, and to illustrate the use of these tables as applied to actual experimental data.

The function i/S given in Table 10 is somewhat intricate analytically. It may be expressed in terms of the function

$$\frac{S}{\alpha} = -\log(1-x) = t$$

in the form

$$\frac{i}{S} = \frac{1}{12}t^2 + s_2 - \frac{2}{t}s_3 + \sum_{r=1}^{\infty}\frac{1}{r^2}\left(1+\frac{2}{rt}\right)e^{-rt} - \frac{t^2}{4}\frac{1-e^{-t}}{t-1+e^{-t}}$$

(where s_2 and s_3 are the sums of the inverse squares and cubes of the natural numbers), a form useful for the larger values of t; or, in ascending powers of t as

$$\frac{1}{12}t^2 - \frac{1}{144}t^3 + \frac{1}{1080}t^4 + \frac{1}{32400}t^5 - \frac{1}{27216}t^6 - \frac{341}{228{,}614400}t^7 + \cdots,$$

which may be used for values below the range of the table. The argument of the table is

$$\log_{10}\frac{N}{S} = \log_{10}\frac{e^t - 1}{t}.$$

The discussion of the appropriate standard error of α calculated from the two observational values N and S raises questions of some interest, since it would seem possible to adopt either N or S as representing the "size of the sample." The formula given in Section 4 is calculated for simultaneous variations of N and S associated with a fixed degree of sampling activity, this being such that the average values of specimens, N, and of species, S, are taken to be equal to the values observed. A full discussion of this point would have to go rather deep into the foundations of statistical inference.

PART 3. A THEORETICAL DISTRIBUTION FOR THE APPARENT ABUNDANCE OF DIFFERENT SPECIES

By R. A. FISHER

(1) *The Poisson Series and the Negative Binomial distribution*

In biological sampling it has for some time been recognized that if successive, independent, equal samples be taken from homogeneous material, the number of individuals observed in different samples will vary in a definite manner. The distribution of the number observed depends only on one parameter, and may be conveniently expressed in terms of the number *expected*, m, in what is known as the *Poisson Series*, given by the formula

$$e^{-m}\frac{m^n}{n!}. \qquad (1)$$

Here n is the variate representing the number observed in any sample, m is the parameter, the number expected, which is the average value of n, and need not be a whole number. Obviously, m will be proportional to the size of the sample taken, and to the density of organisms in the material sampled. For example, n might stand for the number of bacterial colonies counted on a plate of culture medium, m for the average number in the volume of dilution added to each plate. The formula then gives the probability of obtaining n as the number observed.

The same frequency distribution would be obtained for the numbers of different organisms observed in one sample, if all were equally frequent in the material sampled.

If the material sampled were heterogeneous, or if unequal samples were taken, we should have a mixture of distributions corresponding to different values of m. The same is true of the numbers of different organisms observed in a single sample, if the different species are not equally abundant.

An important extension of the Poisson series is provided by the supposition that the values of m are distributed in a known and simple manner. Since m must be positive, the simplest supposition as to its distribution is that it has the Eulerian form (well known from the distribution of χ^2) such that the element of frequency or probability with which it falls in any infinitesimal range dm is

$$df = \frac{1}{(k-1)!}p^{-k}m^{k-1}e^{-m/p}\,dm. \qquad (2)$$

If we multiply this expression by the probability, set out above, of observing just n organisms, and integrate with respect to m over its whole range from 0 to ∞, we have

$$\int_0^\infty \frac{1}{(k-1)!}p^{-k}m^{k-1}e^{-m/p}e^{-m}\frac{m^n}{n!}\,dm,$$

which, on simplification, is found to have the value

$$\frac{(k+n-1)!}{(k-1)!\,n!}\frac{p^n}{(1+p)^{k+n}}, \qquad (3)$$

which is the probability of observing the number n when sampling from such a heterogeneous population. Since this distribution is related to the negative binomial expansion

$$\left(1-\frac{p}{1+p}\right)^{-k} = \sum_{n=0}^{\infty}\frac{(k+n-1)!}{(k-1)!\,n!}\left(\frac{p}{1+p}\right)^n,$$

it has become known as the *Negative Binomial distribution*. It is a natural extension of the Poisson series, applicable to a somewhat wider class of cases.

The parameter p of the negative binomial distribution is proportional to the size of the sample. The expectation, or mean value of n, is pk. The second parameter k measures in an inverse sense the variability of the different expectations of the component Poisson series. If k is very large these expectations are nearly equal, and the distribution tends to the Poisson form. If heterogeneity is very great k becomes small and approaches its limiting value, zero. This second parameter, k, is thus an intrinsic property of the population sampled.

(2) *The limiting form of the negative binomial, excluding zero observations*

In many of its applications the number n observed in any sample may have all integral values including zero. In its application, however, to the number of representatives of different species obtained in a collection, only frequencies of numbers greater than zero will be observable, since by itself the collection gives no indication of the number of species which are not found in it. Now, the abundance in nature of different species of the same group generally varies very greatly, so that, as I first found in studying Corbet's series of Malayan butterflies, the negative binomial, which often fits such data well, has a value of k so small as to be almost indeterminate in magnitude, or, in other words, indistinguishable from zero. That it is not really zero for collections of wild species follows from the fact that the total number of species, and therefore the total number not included in the collection, is really finite. The real situation, however, in which a large number of species are so rare that their chance of inclusion is small, is well represented by the limiting form taken by the negative binomial distribution, when k tends to zero.

The limiting value $k=0$ cannot occur in cases where the frequency at zero is observable, for the

distribution would then consist wholly of such cases. If, however, we put $k=0$ in expression (3), write x for $p/(p+1)$, so that x stands for a positive number less than unity, varying with the size of the sample, and replace the constant factor $(k-1)!$ in the denominator, by a new constant factor, α, in the numerator, we have an expression for the expected number of species with n individuals, where n now cannot be zero,

$$\frac{\alpha}{n}x^n. \qquad (4)$$

These two relationships enable the series to be fitted to any series of observational data, for if S is the number of species observed, and N the number of individuals, the two equations

$$S=-\alpha \log_e(1-x), \quad N=\alpha x/(1-x),$$

are sufficient to determine the values of α and x. The solution of the equations is, however, troublesome and indirect, so that to facilitate the solution in any particular case I have calculated a table (Table 9) from which, given the common logarithm

Table 9. *Table of* $\log_{10} N/\alpha$ *in terms of* $\log_{10} N/S$, *for solving the equation*

$$S=\alpha \log_e\left(1+\frac{N}{\alpha}\right), \text{ given } S \text{ and } N$$

$\log_{10} N/S$	0	1	2	3	4	5	6	7	8	9
0·4	0·61121	63084	65023	66939	68832	70701	72551	74382	76195	77990
0·5	0·79766	81526	83271	85002	86717	88417	90105	91779	93442	95092
0·6	0·96730	98356	99973	1·01579	03174	04759	06335	07902	09460	11010
0·7	1·12550	14083	15607	17124	18634	20136	21631	23120	24602	26077
0·8	1·27546	29008	30465	31916	33361	34801	36234	37663	39087	40506
0·9	1·41920	43329	44733	46133	47528	48919	50305	51688	53066	54440
1·0	1·55810	57177	58539	59898	61254	62605	63954	65299	66640	67979
1·1	1·69314	70646	71975	73301	74623	75943	77261	78575	79886	81195
1·2	1·82501	83805	85106	86404	87700	88994	90285	91574	92860	94144
1·3	1·95426	96706	97984	99259	2·00532	01804	03073	04340	05605	06869
1·4	2·08130	09389	10647	11902	13156	14409	15659	16908	18155	19400
1·5	2·20644	21886	23126	24365	25602	26838	28072	29305	30536	31766
1·6	2·32994	34221	35446	36670	37893	39114	40334	41553	42770	43986
1·7	2·45201	46414	47627	48838	50048	51256	52464	53670	54875	56079
1·8	2·57282	58484	59684	60884	62083	63280	64476	65672	66866	68059
1·9	2·69252	70443	71633	72822	74011	75198	76385	77570	78755	79939
2·0	2·81121	82303	83484	84664	85843	87022	88199	89376	90552	91727
2·1	2·92901	94075	95247	96419	97590	98760	99930	3·01099	02267	03434
2·2	3·04600	05766	06931	08095	09259	10422	11584	12745	13906	15066
2·3	3·16225	17384	18542	19699	20856	22012	23168	24323	25477	26630
2·4	3·27783	28936	30087	31238	32389	33539	34688	35837	36985	38133
2·5	3·39280	40426	41572	42717	43862	45006	46150	47293	48436	49578
2·6	3·50719	51860	53001	54141	55280	56419	57558	58696	59833	60970
2·7	3·62106	63242	64378	65513	66648	67782	68915	70048	71181	72313
2·8	3·73445	74577	75707	76838	77968	79097	80227	81355	82484	83611
2·9	3·84739	85866	86992	88119	89244	90370	91495	92619	93743	94867
3·0	3·95991	97114	98236	99358	4·00480	01602	02723	03843	04964	06084
3·1	4·07203	08322	09441	10560	11678	12795	13913	15030	16147	17263
3·2	4·18379	19494	20610	21725	22839	23954	25068	26181	27295	28408
3·3	4·29520	30632	31744	32856	33967	35079	36189	37300	38410	39520
3·4	4·40629	41738	42847	43956	45064	46172	47280	48387	49494	50601
3·5	4·51707	52814	53920	55025	56131	57236	58340	59445	60549	61653

The total number of species expected is consequently

$$\sum_{n=1}^{\infty}\frac{\alpha}{n}x^n=-\alpha \log_e(1-x),$$

so that our distribution is related to the algebraic expansion of the logarithm, as the negative binomial distribution is to the binomial expansion. Next, it is clear that the total number of individuals expected is

$$\sum_{n=1}^{\infty}\alpha x^n=\frac{\alpha x}{1-x}.$$

of N/S, we may obtain that of N/α. Five-figure logarithms are advisable, such as those in *Statistical Tables*. If x be eliminated from the two equations, it appears that

$$N=\alpha(e^{S/\alpha}-1), \quad S=\alpha \log_e\left(1+\frac{N}{\alpha}\right),$$

and

$$\frac{N}{S}=(e^{S/\alpha}-1)\div S/\alpha,$$

from which Table 9 has been constructed.

(3) *Fitting the series*

The use of the table is shown, using Williams's extensive data for the Macrolepidoptera at Harpenden (total catch for four years). Symbols + and − are used to indicate numbers to be added and subtracted respectively.

Symbol	Number	Common logarithm
S	240	−2·38021
N	15609	+4·19338
N/S	—	1·81317
From the table	$\log(N/S)$	$\log(N/\alpha)$
	−1·81	−2·58484
	+1·82	+2·59684
Difference	0·01	0·01200
Proportional parts	0·00317	0·00380
	1·81317	2·58864
Then	Number	Common logarithm
N/α	—	−2·58864
N	—	+4·19338
α	40·248	1·60474

For constructing the distribution we should then calculate

$$x=\frac{N}{N+\alpha}=\frac{15609}{15649\cdot248}=0\cdot9974281.$$

The quantity α is independent of the size of sample, and is proportional to the number of species of the group considered, at any chosen level of abundance, relative to the means of capture employed. Values of α from different samples or obtained by different methods of capture may therefore be compared as a measure of richness in species. To this end we shall need to know the sampling errors by which an estimate of α may be affected.

(4) *Variation in parallel samples*

Whatever method of capture may be employed, it is to be expected that a given amount of activity devoted to it, e.g. a given number of hours exposure of a light-trap, or a given volume of sea water passed through a plankton filter, will yield on different occasions different numbers of individuals and of species, and, consequently, varying estimates of α. The amount of variation of these kinds attributable to chance must form the basis of all conclusions as to whether variations beyond chance have occurred in the circumstances in which two or more samples were made.

In strictly parallel samples, i.e. equivalent sampling processes applied to homogeneous material, the numbers caught of each individual species will be distributed in a Poisson series, and it easily follows that the same is true of the aggregate number, N, of all species. Since N is a large number of hundreds or thousands, this is equivalent to N being normally distributed with a variance equal to its mean, so that to any observed value N we may attach a standard error (of random sampling) equal to $\pm\sqrt{N}$.

For the variation of S we must obtain the distribution of species according to the number m expected in the sample; modifying expression (2) in the same way as (3) has been modified, this is found to be

$$\alpha e^{-\alpha m/N}dm/m. \qquad (5)$$

The probability of missing any species is e^{-m}, so that the contribution to the sampling variance of S due to any one species being sometimes observed and sometimes not, is

$$e^{-m}(1-e^{-m}).$$

Multiplying this by the frequencies in (5) and integrating over all values of m, we have

$$\alpha\int_0^\infty e^{-m(N+\alpha)/N}\left(1-\frac{m}{2}+\frac{m^2}{6}-\ldots\right)dm=\alpha\log_e\left(\frac{2N+\alpha}{N+\alpha}\right),$$

which is the sampling variance of S. For large samples this is approximately $(0\cdot6931)\,\alpha$.

Variations of S and N in parallel samples are not, however, independent. When present, a species must contribute on the average $m/(1-e^{-m})$ individuals, which exceeds the expectation in all samples by

$$\frac{me^{-m}}{1-e^{-m}},$$

and as the frequency of occurrence is $1-e^{-m}$, each species must contribute $m.e^{-m}$ to the covariance of S and N. The covariance is thus found to be

$$\frac{\alpha N}{N+\alpha}.$$

From these three values it is possible by standard methods to find the sampling variance of S in samples having a given number of specimens N, which is

$$V(S),\ \text{given } N,=\alpha\log_e\frac{2N+\alpha}{N+\alpha}-\frac{\alpha^2N}{(N+\alpha)^2}$$

and, the variance of α,

$$V(\alpha)=\frac{\alpha^3\left\{(N+\alpha)^2\log_e\frac{2N+\alpha}{N+\alpha}-\alpha N\right\}}{(SN+S\alpha-N\alpha)^2}.$$

We may, therefore, complete the example of the last section by calculating the standard error of α. Using the values obtained, the variance comes to 1·1251, of which the square root is 1·0607.

The estimate obtained for α, 40·248, has, therefore, a standard error of 1·0607, available for comparison with like estimates.

(5) *Test of adequacy of the limiting distribution*

From the manner in which the distribution has been developed it appears that we never have theoretical grounds for supposing that k is actually zero;

but, on the contrary, must generally suppose that in reality it has a finite, though perhaps a very small, value. Our reasons for supposing this small value to be negligible must always be derived from the observations themselves. It is, therefore, essential to be able to test any body of data in respect to the possibility that in reality some value of k differing significantly from zero might fit the data better than the value zero actually assumed.

The most sensitive index or score by which any departure of the series of frequencies observed from those expected can be recognized, is found by the general principles of the Theory of Estimation, as, for example, in the author's *Statistical Methods for Research Workers*, to be

$$S\left\{a_n\left(1+\tfrac{1}{2}+\tfrac{1}{3}+\ldots+\frac{1}{n-1}\right)\right\},$$

when a_n is a number of species observed with n individuals in each. If the values of a_n conformed accurately with expectation, the total score would be equal to

$$\frac{S^2}{2\alpha}.$$

If, on the contrary, the series were better fitted by a negative binomial with a value of k differing from zero, we should expect the difference

$$S\left\{a_n\left(1+\tfrac{1}{2}+\tfrac{1}{3}+\ldots+\frac{1}{n-1}\right)\right\}-\frac{S^2}{2\alpha}$$

to show a positive discrepancy.

Applying this test to Williams's distribution for 240 species of Macrolepidoptera, one finds, after a somewhat tedious calculation,

$S\left\{a_n\left(1+\tfrac{1}{2}+\tfrac{1}{3}+\ldots+\frac{1}{n-1}\right)\right\}$	724·86
$\frac{S^2}{2\alpha}$	715·57
Difference	+9·29

The series, therefore, shows a deviation in the direction to be expected for the negative binomial, though apparently quite a small one. In order to test the significance of such discrepancies, I give in Table 10, for the same range of observable values of the average number of specimens in each species N/S, the values of i/S, where i is the quantity of information, in respect of the value of k, which the data supply.

Table 10. *The amount of information respecting k, supposed small, according to the numbers of individuals (N) and species (S) observed*

$\log_{10} N/S$	i/S	$\log_{10} N/S$	i/S
0·4	0·1971		
0·5	0·2882		
0·6	0·3914	2·1	3·1047
0·7	0·5054	2·2	3·3606
0·8	0·6295	2·3	3·6260
0·9	0·7639	2·4	3·9009
1·0	0·9076	2·5	4·1854
1·1	1·0608	2·6	4·4791
1·2	1·2232	2·7	4·7825
1·3	1·3950	2·8	5·0954
1·4	1·5762	2·9	5·4178
1·5	1·7665	3·0	5·7498
1·6	1·9661	3·1	6·0912
1·7	2·1751	3·2	6·4421
1·8	2·3934	3·3	6·8026
1·9	2·6211	3·4	7·1726
2·0	2·8582	3·5	7·5521

Entering the table with our value 1·81317 for $\log_{10} N/S$ we have $i/S=2{\cdot}4656$, or $i=591{\cdot}7$. This quantity may now be used for two purposes. In the first place it is the sampling variance of the discrepancy observed, so that, taking its square root, the standard error is found to be 24·33. This suffices to test the significance of the discrepancy, since $9{\cdot}29 \pm 24{\cdot}33$ is clearly insignificant.

If, on the contrary, a significant discrepancy had been found, an estimate of the value of k required to give a good fit to the data could be made by dividing the discrepancy by i. In fact

$$\frac{9{\cdot}29}{591{\cdot}7}=0{\cdot}016$$

would have been the value of k indicated by the data, if any value other than zero had been required.

REFERENCES

Fisher, R. A. & Yates, F. (1943). 'Statistical tables for biological, agricultural and medical research' (2nd ed.). Edinburgh.

Fisher, R. A. (1941). 'Statistical methods for research workers' (8th ed.). Edinburgh.

SUMMARY

Part 3. A theoretical distribution is developed which appears to be suitable for the frequencies with which different species occur in a random collection, in the common case in which many species are so rare that their chance of inclusion is small.

The relationships of the new distribution with the negative binomial and the Poisson series are established.

Numerical processes are exhibited for fitting the series to observations containing given numbers of species and individuals, and for estimating the parameter α representing the richness in species of the material sampled; secondly, for calculating the standard error of α, and thirdly, for testing whether the series exhibits a significant deviation from the limiting form used.

Special tables are presented for facilitating these calculations.