A note on some asymptotic properties of the logarithmic series distribution

BY J. C. GOWER

Rothamsted Experimental Station

SUMMARY

Formulae are given to calculate

$$S(x, R) = \sum_{r=R+1}^{\infty} \frac{x^r}{r},$$

the remainder term in the logarithmic series, which are valid for high values of R and values of x close to unity. These are of value when dealing with extreme values of the parameters appearing in the logarithmic series distribution such as have arisen in C. B. Williams's recent study of the possible distribution of the number per species for all the insects in the world. Bounds are given for the errors involved when making the recommended approximations. The calculation of the median of the distribution is also discussed.

INTRODUCTION

The Logarithmic Series Distribution was introduced in an article by Fisher, Corbett & Williams (1943) who fitted it to the numbers of individuals per species in samples of Macro-Lepidoptera caught in a light trap at Rothamsted. Since then a number of papers, notably those of C. B. Williams, have used this distribution to describe data of diverse kinds. Recently Williams (1960) has investigated the problem of assessing the possible distribution of the number per species of all the insects in the world. Attention is focused on the number of species with more than a specified number of individuals. When this specified number is very large and the logarithmic distribution is used, a knowledge of its asymptotic properties is needed. Williams (1960) remarks that 'this is a complex problem which requires further investigation'. In this article some asymptotic properties of the logarithmic distribution will be listed and some indication is given of the range of values of the parameters of this distribution for which the results are reasonably accurate. Biologists should note these restrictions, since if the formulae are applied outside the range for which they are designed, inaccurate results may be obtained.

Using the standard notation, suppose we have N individuals distributed in S different species. Then Fisher showed on quite plausible assumptions (Fisher *et al.* 1943) that the expected number of species with i individuals would be proportional to x^i/i where x is a parameter, having a value between 0 and 1, which depends on N and S. If the factor of proportionality is α then the expected numbers with 1, 2, 3, etc., individuals per species are αx, $\alpha x^2/2$, $\alpha x^3/3$, ... These are the individual terms in the series expansion of $-\alpha \log(1-x)$ and it is for this reason that the series is known as the logarithmic series. The total number of individuals is $\alpha x + \alpha x^2 + \alpha x^3 + \ldots = \alpha x/(1-x)$. To estimate α and x we thus have two equations

$$\left.\begin{aligned} S &= -\alpha \log(1-x), \\ N &= \frac{\alpha x}{1-x}. \end{aligned}\right\} \tag{1}$$

It has often been found that, using formulae (1) to estimate α and x for a set of data, the logarithmic series gives a good fit to the observations. Williams (1960) postulates that it might not be unreasonable to use this series to fit to the number of insects and insect species in the world, although a better fit might be obtained by using a truncated lognormal distribution. As rough estimates he takes $N = 10^{18}$ and $S = 3 \times 10^6$ and with these values calculates $x = 1 - 1{\cdot}002 \times 10^{-13}$ and $\alpha = 1{\cdot}002 \times 10^5$. He then asks how many species there are with more than R individuals. This quantity is

$$\alpha S(x, R) = \sum_{r=R+1}^{\infty} \alpha x^r / r.$$

The remainder of this article is concerned with the numerical evaluation of $S(x, R)$ particularly for large values of R.

EVALUATION OF THE REMAINDER $S(x, R)$

An exact formula for the remainder term, true for all values of x and R is

$$S(x, R) = \log(1-x) - \sum_{1}^{R} 1/n - \int_0^{1-x} \frac{(1-\phi)^R - 1}{\phi}\, d\phi. \tag{2}$$

To evaluate this formula for large values of R we deal first with the term $\sum_1^R 1/n$. This has been tabulated by Glover (1930) for $R = 2(1)450$. If these tables are not available or if larger values of R are of interest recourse can be had to the well-known formula

$$\sum_1^R 1/n = \gamma + \tfrac{1}{2}R + \log R - U_R, \tag{3}$$

where $\gamma = 0{\cdot}5772156649$ is Euler's constant and

$$U_R = \frac{B_1}{2R^2} - \frac{B_2}{4R^4} + \frac{B_3}{6R^6} - \ldots,$$

the B_i being Bernoulli numbers. The series U_R rapidly approaches zero as R increases, specimen values being given in Table 1.

Table 1

R	U_R
10	0·00083 25039
50	·00003 33320
100	·00000 83333
200	·00000 20833
500	·00000 03333
1,000	·00000 00833
2,000	·00000 00208
5,000	·00000 00033
10,000	·00000 00008

It is thus permissible to replace $\sum_1^R 1/n$ by

$$\gamma + \tfrac{1}{2}R + \log R. \tag{4}$$

If $R > 100$ the result is at most one figure out in the fifth decimal place.

Returning to (2) we now try to find an expression for the integral term more amenable to calculation. A well-known inequality is

$$e^{-t} \geqslant (1 - t/n)^n \geqslant e^{-t} - t^2/(2n) \quad \text{if} \quad 0 \leqslant t \leqslant n.$$

Writing $n = R$ and $t = \phi R$, subtracting one, dividing each expression by ϕ, integrating between 0 and $(1-x)$ and noting that

$$\int_0^{1-x} \frac{e^{-R\phi} - 1}{\phi}\, d\phi = -\mathrm{Ei}[-R(1-x)] + \log R(1-x) + \gamma \tag{5}$$

we find that

$$-\mathrm{Ei}[-R(1-x)] + \log[R(1-x)] + \gamma \geqslant \int_0^{1-x} \frac{(1-\phi)^R - 1}{\phi}\, d\phi$$
$$\geqslant -\mathrm{Ei}[-R(1-x)] + \log[R(1-x)] + \gamma - \tfrac{1}{2}R(1-x)^2.$$

Here $-\mathrm{Ei}(-u)$ is the exponential integral and has been tabulated by several authors, the most extensive tabulation being that of the New York W.P.A. (1940) where $u = 0(0{\cdot}001)10(0{\cdot}1)15$. Results for some higher values of u are given by Akahiro (1929) where $u = 20(0{\cdot}02)50$. We may thus write approximately

$$\int_0^{1-x} \frac{(1-\phi)^R - 1}{\phi}\, d\phi = -\mathrm{Ei}[R(1-x)] + \log R(1-x) + \gamma, \tag{6}$$

the error being less than $\frac{1}{2}R(1-x)^2$. Combining (2), (3) and (6)

$$S(x, R) = -\mathrm{Ei}[-R(1-x)] - \frac{1}{2R} + U_R + V, \tag{7}$$

where V is an error term which is less than $\frac{1}{4}R(1-x)^2$. A further simplification can be made when $R(1-x)$ is small. In this case

$$\int_0^{1-x} \frac{(1-\phi)^R - 1}{\phi}\, d\phi = -R(1-x) \tag{8}$$

with an error less than $\frac{1}{4}R^2(1-x)^2$. Thus we have

$$S(x, R) = -\log(1-x) - \sum_1^R 1/n + R(1-x) + RV. \tag{9}$$

An alternative approach, due to P. M. Grundy (unpublished work) is to express $\int_R^\infty \frac{X^u}{u}\,du$ in terms of its Euler–Maclaurin expansion (Jeffreys & Jeffreys, 1950, pp. 278–83). This gives

$$S(x, R-1) = \mathrm{Ei}\,[R\log x] + x^R/2R \tag{10}$$

with an error less than $(1 - R\log x)\,x^R/(12R^2)$. At first sight this seems a neater result than those obtained above, and in fact for many values of x and R it is. Unfortunately, for very large values of R and x near to unity, x^R must for ease of computation be replaced by $e^{-R(1-x)}$ and an investigation of the error involved in this approximation would invoke inequalities such as the one used previously. In fact, making this approximation, (10) can be shown to be equivalent to (7).

The use of the formulae

The formulae appear at first sight to be rather complex, but in fact little difficulty should be found in using them. A summary of the situations when the different formulae should be used is given below.

(1) $R^2(1-x)^2$ *is small.* Use formula (9). The error is less than $\frac{1}{3}R^2(1-x)^3$. If tables of $\sum_1^R 1/n$ are not available, or if R is too large formula (3) may be used to evaluate $\sum_1^R 1/n$.

(2) $R(1-x)^2$ *is small.* If $R(1-x)$ is appreciable but $R(1-x)^2$ is small use formula (7). The error is less than $\frac{1}{3}R(1-x)^3$.

(3) If x is not too small use Grundy's formula (10).

These formulae should be sufficient for most calculations.

Table 2. *Calculation of the remainder term* $S(x, R)$ *for* 3×10^6 $(= S)$ *species and* $x = 1 - 1{\cdot}002 \times 10^{-13}$

R	$S(x, R)$	$100\left\{1 + \frac{S(x,R)}{\log(1-x)}\right\}$	Notes	Method
1	28·932	3·3	—	Direct summation
10	27·003	9·8	Formulae (9) and (2) (with the term $1/(2R)$) gives 27·002	
10^2	24·744	17·3	Direct summation gives the same results	Formulae (9) and (2), including the term $1/(2R)$
10^3	22·446	25·0		
10^4	20·144	32·7	—	Formulae (9) and (2) (the term $1/(2R)$ is negligible)
10^5	17·842	40·4	—	
10^6	15·539	48·1	—	
10^7	13·236	55·8	—	
10^8	10·934	63·5	—	
10^9	8·631	71·2	—	
10^{10}	6·330	78·9	—	
10^{11}	4·038	86·5	These values agree with those obtained by formulae (2) and (9)	Formula (7)
10^{12}	1·823	93·9		
10^{13}	0·219	99·3	—	
10^{14}	0·000	100·0	—	

As an example Williams (1960) suggested values of $N = 10^{13}$ and $S = 3 \times 10^6$ for the number of insects and insect species in the world and calculates $x = 1 - 1{\cdot}002\,(10^{-13})$, $\alpha = 1{\cdot}002\,(10^5)$. With these figures and the formulae derived above Table 2 is produced. This agrees substantially with Williams's Table 4 (see the columns for $S = 3 \times 10^6$) and extends his table for the higher values of R. If the other two columns of Williams's Table 4 are extended (i.e. $S = 2 \times 10^6$ and $S = 5 \times 10^6$), similar sets of results are obtained and the main conclusion, that the percentage number of species with more than a specified number of individuals per species remains fairly steady over the range of values of S considered, remains true for the higher values of R.

The median

The median of the logarithmic series is defined as the value of R satisfying the equation

$$S(x, R) = -\tfrac{1}{2}\log(1-x). \tag{11}$$

If R is large (> 100) and $R(1-x)$ is small, then (9) becomes $S(x, R) = -\log R(1-x) - \log R - \gamma$ so that the median is given by $-\frac{1}{2}\log(1-x) = \log R + \gamma$, i.e.

$$R = \frac{e^{-\gamma}}{\sqrt{(1-x)}} = \frac{0{\cdot}56146}{\sqrt{(1-x)}}. \tag{12}$$

This, apart from a term 0·81524, is the formula given on page 144 of Williams (1960) and attributed to P. M. Grundy. The 0·81524 given should there occur as an additive term and not a multiplicative one as printed.

A further stage of approximation adds in the Grundy terms $(\frac{1}{2} + e^{-2\gamma})$, but for high values of S this is of course quite trivial. In many ways the best approach appears to be to produce a table similar to Table 2 above and then interpolate to find the value of $R-1$ corresponding to 50 %. For Table 2 this would occur between $R = 10^6$ and $R = 10^7$ and where the curve of $S(x, R)$ is linear relative to a log-scale of R. Noting that $-\log(1-x) = 29{\cdot}932$, linear interpolation gives

$$15{\cdot}539 - 13{\cdot}236 = \frac{15{\cdot}539 - \frac{1}{2}(29{\cdot}932)}{7 - \log R},$$

$\log R = 6{\cdot}2488$, $R = 1773000$, the value obtained by formula (12). If the median occurs on the curved portion of the $S(x, R)$ curve then some more elaborate form of inverse interpolation using, say, four adjacent points on the $S(x, R)$ curve, will have to be used.

References

Fisher, R. A., Corbet, A. S. & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.

Williams, C. B. (1960). The range and pattern of insect abundance. *Amer. Nat.* **94**, 137–51.

Glover, J. W. (1930). *Tables of Applied Mathematics in Finance, Insurance and Statistics.* Ann Arbor, Michigan: Wahr.

New York W.P.A. (1940). *Tables of Sine, Cosine and Exponential Integrals*, **1** and **2**.

Akahiro, T. (1929). *Sci. Pap. Inst. Phys. Chem. Res. Tokyo.* Table no. 3, pp. 181–215.

Jeffreys, H. & Jeffreys, B. S. (1950). *Mathematical Physics* (second edition). Cambridge University Press.

On a property of balanced designs

By M. Atiqullah

Birkbeck College, London

1. Introduction and summary

Special classes of balanced designs are well known and widely used. In this paper, a necessary and sufficient condition for a general class of connected designs to be balanced is derived. This is a natural extension of a result of Tocher (1952) and of Thompson (1956), and it appears to be simpler than the generalization given by Rao (1958). A simple expression for calculating the efficiency factor for a connected balanced design is obtained.

Fisher (1940) established that $b \geqslant v$ for a balanced incomplete block design with v treatments and b blocks. This inequality is shown to be true for a wider class of designs, similar to the balanced incomplete block designs but with blocks of different sizes.

2. Notation and preliminary results

Consider v treatments arranged in b blocks in a design whose incidence matrix is $\mathbf{N} = (n_{ij})$, where n_{ij} denotes the number of experimental units in the ith block getting the jth treatment. The ith block is of size k_i ($i = 1, 2, \ldots, b$) and the jth treatmen tis replicated r_j times ($j = 1, 2, \ldots, v$). These may be