THE GEOMETRIC, LOGARITHMIC AND DISCRETE PARETO FORMS OF SERIES

By P. G. MOORE, Ph.D., A.I.A.

University College, London

Although the Poisson distribution is by far the best known of the one-parameter form of discrete distribution and has been widely applied in practice, there are several other such distributions in existence. In this note it is proposed to discuss three of these distributions, namely, the geometric, logarithmic and discrete Pareto forms, and compare their properties. Some illustrations of the use of the series will be made.

2. The geometric series takes the form

$$p_r = p^{r-1}(1-p) \quad (r = 1, 2, \ldots), \tag{1}$$

where p_r is the probability of just r successes or events. Thus each probability is a fixed multiple of the previous probability, and since p must be less than unity for the sum of the p_r's to be unity, it follows that every term is smaller than the previous term. Hence the first term is always the largest, giving a J-shaped distribution. Suppose that in a sample of N observations there are n_r observations having the value r, then the likelihood of the sample arising is proportional to

$$P = (1-p)^N p^{n_1}(p^2)^{n_2} \ldots (p^j)^{n_j+1} \ldots,$$

whence

$$\log P = N \log (1-p) + (\log p) \sum_{r=1}^{\infty} r n_{r+1}. \tag{2}$$

By differentiating (2) with respect to p and equating to zero we find that the maximum-likelihood solution for p is given by the equation

$$\hat{p} = \frac{\mu_1' - 1}{\mu_1'}, \tag{3}$$

where μ_1' is the first moment of the distribution of the n_j. To find the variance of the estimate we perform a second differentiation of (2) above with respect to p (see, for example, *Johnson & Tetley* (1950), p. 129) leading to

$$N \operatorname{Var}(\hat{p}) = \frac{\mu_1' - 1}{\mu_1'^2}. \tag{4}$$

The characteristic property of the geometric form of series is that the probability of there being at least $r+1$ events, given that there are at least r, is exactly the same whatever the value of r. This may be easily demonstrated, since the probability of there being at least $r+1$ events, given there are at least r, is

$$\frac{p_{r+1} + p_{r+2} + \ldots}{p_r + p_{r+1} + p_{r+2} + \ldots} = \frac{p + p^2 + p^3 + \ldots}{1 + p + p^2 + p^3 + \ldots} = p, \text{ i.e. independent of } r.$$

Another property that should be noted is that since

$$\log p_r = (r-1) \log p + \log (1-p),$$

there is a linear relationship between r and $\log p_r$ and hence a graph of r against $\log p_r$ would be a straight line. This gives a quick method of seeing in practice whether a given set of data may be graduated by a geometric series, since first differences of $\log n_r$ can be quickly calculated and should be approximately constant.

The series has only infrequently been used in practical cases. For example, the interval between two successive zeros in a table of random numbers, 0–9, should follow a geometric series, since

$$\text{probability that interval is of length } r=(\tfrac{9}{10})^{r-1}\tfrac{1}{10} \quad (r\geqslant 1).$$

A similar type of situation obtains if we consider the runs of red or black on a roulette wheel. Dufrenoy (1938) considered the distribution of the number of papers published by biologists in any one year. He suggested that the geometric series might give a satisfactory fit and tried it using a trial and error value for p. His fit was only moderate, and the use of maximum likelihood to estimate p does not greatly improve the fit.

3. The logarithmic series takes the form

$$p_r=\frac{p_1 x^{r-1}}{r} \quad (r=1, 2, 3, \ldots), \tag{5}$$

and by making $\sum_{r=1}^{\infty} p_r=1$ we obtain $p_1=-x/\log_e(1-x)$. The likelihood of a sample of N observations of which n_r have the value r is therefore proportional to

$$P=\left\{\frac{-x}{\log_e(1-x)}\right\}^N\left(\frac{x}{2}\right)^{n_2}\left(\frac{x^2}{3}\right)^{n_3}\cdots\left(\frac{x^{r-1}}{r}\right)^{n_r}\cdots.$$

By taking logarithms and equating to zero the differential with respect to x we find

$$\mu_1'=\frac{-x}{(1-x)\log_e(1-x)}, \tag{6}$$

where μ_1' is the mean or first moment of the n_r. Thus maximum likelihood leads to the equating of the observed and theoretical first moments. In fitting an observed set of data with this series it is necessary to solve (6) above for x. This is apt to be troublesome and Table 1 is designed to enable this to be done directly by linear interpolation. The table was constructed by first making a table of μ_1' for a suitable range of values of x and then inversely interpolating in that table. Since p_1 is equal to $(1-x)\mu_1'$ it can be found as soon as x has been obtained and thus the expected frequencies may be written down. Fisher, Corbett & Williams (1943) have given a table for obtaining x in somewhat different form requiring the use of a table of logarithms as well. The variance of the maximum-likelihood estimate of x is given by

$$N\ \text{Var}\,(\hat{x})=\frac{x^2}{\mu_1'}\left[\frac{1}{1-\mu_1'\{1+\log_e(1-x)\}}\right], \tag{7}$$

and in Table 2 some specimen values are given for various values of x. For the logarithmic series the probability of there being at least $r+1$ events given that there are at least r is

$$\frac{p_{r+1}+p_{r+2}+p_{r+3}+\ldots}{p_r+p_{r+1}+p_{r+2}+\ldots}=\frac{\dfrac{x^r}{r+1}+\dfrac{x^{r+1}}{r+2}+\ldots}{\dfrac{x^{r-1}}{r}+\dfrac{x^r}{r+1}+\ldots}=1-\frac{1}{1+\dfrac{r}{r+1}x+\dfrac{r}{r+2}x^2+\ldots}$$

after some simplification. As r increases the ratio $r/r+s$ increases for fixed s, and hence the denominator of the second term increases which implies that the whole expression increases. Thus the logarithmic series differs from the geometric series in that the probability of at least $r+1$ events occurring given that at least r have occurred increases as r increases and does not remain constant as before.

Table 1. μ_1' against x

μ_1'	x	μ_1'	x	μ_1'	x	μ_1'	x	μ_1'	x
1·08	0·1411	1·26	0·3596	1·44	0·4983	1·62	0·5926	1·80	0·6602
1·09	0·1564	1·27	0·3688	1·45	0·5045	1·63	0·5970	1·81	0·6633
1·10	0·1712	1·28	0·3781	1·46	0·5105	1·64	0·6012	1·82	0·6665
1·11	0·1855	1·29	0·3869	1·47	0·5164	1·65	0·6054	1·83	0·6696
1·12	0·1994	1·30	0·3956	1·48	0·5223	1·66	0·6095	1·84	0·6726
1·13	0·2130	1·31	0·4041	1·49	0·5280	1·67	0·6135	1·85	0·6755
1·14	0·2238	1·32	0·4124	1·50	0·5336	1·68	0·6175	1·86	0·6785
1·15	0·2395	1·33	0·4205	1·51	0·5391	1·69	0·6214	1·87	0·6814
1·16	0·2515	1·34	0·4284	1·52	0·5444	1·70	0·6252	1·88	0·6842
1·17	0·2637	1·35	0·4361	1·53	0·5497	1·71	0·6290	1·89	0·6870
1·18	0·2755	1·36	0·4436	1·54	0·5548	1·72	0·6327	1·90	0·6898
1·19	0·2870	1·37	0·4510	1·55	0·5599	1·73	0·6363	1·91	0·6925
1·20	0·2984	1·38	0·4583	1·56	0·5648	1·74	0·6400	1·92	0·6952
1·21	0·3090	1·39	0·4653	1·57	0·5697	1·75	0·6434	1·93	0·6979
1·22	0·3198	1·40	0·4722	1·58	0·5744	1·76	0·6469	1·94	0·7005
1·23	0·3300	1·41	0·4789	1·59	0·5791	1·77	0·6503	1·95	0·7030
1·24	0·3402	1·42	0·4855	1·60	0·5837	1·78	0·6536	1·96	0·7055
1·25	0·3499	1·43	0·4920	1·61	0·5882	1·79	0·6569	1·97	0·7080

Table 2. Variance of estimate of x

x	N Var (x)	x	N Var (x)	x	N Var (x)	x	N Var (x)
0·14	0·217525	0·28	0·322934	0·42	0·336132	0·56	0·280003
0·16	0·238982	0·30	0·329971	0·44	0·331780	0·58	0·267808
0·18	0·258285	0·32	0·335173	0·46	0·326077	0·60	0·254834
0·20	0·275453	0·34	0·338614	0·48	0·319106	0·62	0·241152
0·22	0·290334	0·36	0·340366	0·50	0·310933	0·64	0·226844
0·24	0·303189	0·38	0·340471	0·52	0·301639	0·66	0·212034
0·26	0·314041	0·40	0·339056	0·54	0·291303	0·68	0·196768

The logarithmic distribution has been used to describe the number of species represented in random collections of plants as well as for the distribution of the number of papers published by scientists (see, for example, Williams (1944)).

4. The discrete form of Pareto's so-called law was first used by Seal (1947) and can be written as

$$p_r=\frac{r^{-\beta}}{\zeta(\beta)} \quad (r=1, 2, 3, \ldots; \beta>1), \tag{8}$$

where $\zeta(\beta)$ is the Riemann Zeta function, since we must have that $\sum_{r=1}^{\infty} r^{-\beta}=\zeta(\beta)$ in order to have the sum of the probabilities equal to unity. In Seal's paper of

1947 he fits the distribution to an observed series by taking the ratio p_1/p_2 which is equal to 2^β as a method of estimating β. In a further paper (Seal, 1952) he uses maximum-likelihood methods, which lead to the equation

$$-\frac{\zeta'(\beta)}{\zeta(\beta)}=\frac{1}{N}\sum_{r=1}^{\infty} n_r \log r. \tag{9}$$

By taking the second differential of the natural logarithm of the likelihood we obtain for the variance of the estimate of β

$$N \text{ Var } (\hat{\beta})=\frac{\{\zeta(\beta)\}^2}{\zeta(\beta)\,\zeta''(\beta)-\{\zeta'(\beta)\}^2}. \tag{10}$$

Equation (9) can be solved numerically by means of a table of $\zeta'(\beta)/\zeta(\beta)$ which has been provided by Walther (1926). For the variance some values of the function in (10) are tabulated in Table 3. Since no table of $\zeta''(\beta)$ has been

Table 3. Variance of estimate of β

β	N Var (β)	β	N Var (β)	β	N Var (β)
2·5	2·914	4·0	17·852	5·5	69·178
3·0	5·832	4·5	28·863	6·0	104·060
3·5	10·552	5·0	45·201		

found the values were obtained by numerical differentiation of a table of $\zeta(\beta)$. It is thought that the table is accurate to the number of significant figures that are given. An alternative method of estimating β (referred to by Seal (1952)) would be to equate the first moments of the observed and theoretical distributions. For (8) we find

$$\mu_1'=\frac{\zeta(\beta-1)}{\zeta(\beta)}, \tag{11}$$

and a table of (11) has been constructed by first calculating μ_1' for various values of β and then inversely interpolating in it to produce a table of β against μ_1'. This is shown as Table 4 and should enable β to be found directly given the mean of the observed distribution. The range of values of μ_1' shown should cover any possible cases met with in practice. It will be noticed that β changes much more rapidly with low values of μ_1' than it does with higher values of μ_1'. Linear interpolation may be used for μ_1' greater than 1·15, but second order should be used below that figure.

Considering once again the probability of at least $r+1$ events occurring given that at least r have occurred we find that

$$\frac{p_{r+1}+p_{r+2}+\dots}{p_r+p_{r+1}+\dots}=1-\frac{1}{1+\left(1+\frac{1}{r}\right)^{-\beta}+\left(1+\frac{2}{r}\right)^{-\beta}+\dots}.$$

Since $(1+1/r)^{-\beta}$ increases as r increases, β being positive, we have that this probability increases, that is to say, the occurrence of r events makes an $(r+1)$th the more likely the larger the value of r.

Table 4. β in terms of μ_1'

μ_1'	β	μ_1'	β	μ_1'	β	μ_1'	β	μ_1'	β
1·01	6·848	1·21	3·419	1·41	2·931	1·61	2·702	1·81	2·566
1·02	5·952	1·22	3·381	1·42	2·915	1·62	2·693	1·82	2·560
1·03	5·449	1·23	3·345	1·43	2·901	1·63	2·685	1·83	2·555
1·04	5·105	1·24	3·311	1·44	2·887	1·64	2·677	1·84	2·550
1·05	4·847	1·25	3·279	1·45	2·873	1·65	2·669	1·85	2·545
1·06	4·642	1·26	3·249	1·46	2·860	1·66	2·662	1·86	2·539
1·07	4·473	1·27	3·220	1·47	2·847	1·67	2·654	1·87	2·535
1·08	4·331	1·28	3·193	1·48	2·834	1·68	2·647	1·88	2·530
1·09	4·208	1·29	3·167	1·49	2·822	1·69	2·640	1·89	2·525
1·10	4·100	1·30	3·143	1·50	2·811	1·70	2·633	1·90	2·520
1·11	4·006	1·31	3·119	1·51	2·799	1·71	2·626	1·91	2·516
1·12	3·921	1·32	3·097	1·52	2·788	1·72	2·619	1·92	2·511
1·13	3·844	1·33	3·075	1·53	2·778	1·73	2·613	1·93	2·507
1·14	3·775	1·34	3·054	1·54	2·767	1·74	2·607	1·94	2·503
1·15	3·711	1·35	3·034	1·55	2·757	1·75	2·600	1·95	2·499
1·16	3·653	1·36	3·015	1·56	2·747	1·76	2·594	1·96	2·495
1·17	3·599	1·37	2·997	1·57	2·737	1·77	2·589	1·97	2·491
1·18	3·549	1·38	2·980	1·58	2·728	1·78	2·583	1·98	2·487
1·19	3·503	1·39	2·963	1·59	2·719	1·79	2·577	1·99	2·483
1·20	3·459	1·40	2·946	1·60	2·710	1·80	2·571	2·00	2·479

5. These distributions may be compared from a number of points of view. We note first of all that the ratio of p_{n+1}/p_n varies both in form and in its limit as n tends to infinity:

Poisson (for comparison)	$p_{n+1}/p_n=\dfrac{\lambda}{n+1}$	tends to 0 $(\lambda>0)$
Geometric	$p_{n+1}/p_n=p$	i.e. constant for all n $(p<1)$
Logarithmic	$p_{n+1}/p_n=\dfrac{n}{n+1}x$	tends to x $(x<1)$
Pareto	$p_{n+1}/p_n=\left(\dfrac{n}{n+1}\right)^{\beta}$	tends to 1 $(\beta>1)$

These comparisons show great differences. Whilst with the Poisson series each term rapidly gets smaller than the previous one, with the geometric series the ratio remains constant but always less than unity. Further, with the logarithmic series the ratio slowly increases from $\frac{1}{2}x$ to x but is always less than unity, whilst with the discrete Pareto form of series it increases from $2^{-\beta}$ to 1 in the limit. This seems to indicate that the length of the 'tail' increases as we go from the Poisson series through the geometric and logarithmic series to the Pareto form of series.

We next note that the probability of $r+1$ or more events given that there are r events varies as follows when r increases:

Poisson	Decreases	Logarithmic	Increases
Geometric	Constant, equal to p	Pareto	Increases

The decrease for the Poisson is to be expected if we remember that the Poisson distribution arises as the limiting form of the binomial distribution $(q+p)^n$ as

n becomes very large and p is small. The geometric series is apparently the intermediate form between the Poisson on the one hand and the logarithmic and Pareto series on the other. The Pareto is very much at the extreme. It has a very long tail and the higher moments are often infinite. For example, we find that

$$\mu_2 = \frac{\zeta(\beta-2)}{\zeta(\beta)} - \left(\frac{\zeta(\beta-1)}{\zeta(\beta)}\right)^2, \tag{12}$$

and thus unless β is greater than 3 the second moment becomes infinite. Of course in practice this is overcome, since a certain amount of truncation at the upper tail of the distribution brings about finite moments. This length of 'tail' must, however, affect the value of the first term, and all the earlier terms, of any of the series. We note as well that the rate of reduction of the terms differs most markedly, and whereas the ratio of succeeding terms remains constant for a geometric distribution it decreases for the Pareto distribution as is shown by the following figures:

Mean of distribution	1·2	1·4	1·6	1·8
Geometric series ratio of r to $(r+1)$th term	6·0	3·5	2·7	2·2
Pareto series ratio of 1st to 2nd term	11·0	7·7	6·5	5·9
Pareto series ratio of 10th to 11th term	1·4	1·3	1·3	1·3

6. We will now briefly illustrate these series by considering some examples. First suppose that a series of articles are being made and the probability that each article possesses some property A is constant and equal to p. It is clear that the probability of a run of r articles occurring between two successive articles possessing property A is $q^{r-1}p$, where $q=1-p$, and hence is a geometric series. In the data given below p is equal to $\frac{1}{2}$, since the property A was whether the length of life of a bulb was above the median value or not. A sequence of bulbs was analysed and the distribution of runs of those above or below the median value was as follows:

Run of length	1	2	3	4	5	>5	Total
Observed number	136	66	26	12	9	5	254
Theoretical number	127	63·5	31·7	15·9	7·9	7·9	253·9

A test of goodness of fit is unnecessary in order to see that there is good agreement between the observed and theoretical frequencies.

For the logarithmic series we will consider the following data concerning the number of cars of different makes and types seen in a period of one hour along an arterial road.

No. of times seen (f)	1	2	3	4	5	6	>6	Total
No. of types seen f times	90	19	9	1	1	3	1	124
Theoretical frequencies	85·0	23·4	8·6	3·6	1·6	0·7	1·1	124

The mean of the observed distribution is 1·5321 and by linear interpolation in Table 1 we find that $x=0·5508$. The theoretical frequencies can now be

calculated and are given above. Without any formal test it can be seen that there is fairly good agreement between observation and theory.

The discrete form of Pareto's law was used by Seal in his papers to graduate the distribution of the duplicate policies on any one life held in an office, the underlying theory being that if the incomes of persons insured are roughly in the form of a Pareto curve the numbers of policies held may also be similarly distributed.

REFERENCES

DUFRENOY, J. (1938). The publishing behaviour of biologists. *Quart. Rev. Biol.* **13**, 207.

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.

JOHNSON, N. L. & TETLEY, H. (1950). *Statistics*, vol. 2. Cambridge University Press.

SEAL, H. L. (1947). A probability distribution of deaths at age x when policies are counted instead of lives. *Skand. AktuarTidskr.* **30**, 18.

SEAL, H. L. (1952). The maximum likelihood fitting of the discrete Pareto law. *J.I.A.* **78**, 115.

WALTHER, A. (1926). Anschauliches zur Riemannschen Zetafunktion. *Acta Math.* **48**, 393.

WILLIAMS, C. B. (1944). The numbers of publications written by biologists. *Ann. Eugen., Lond.*, **12**, 143.