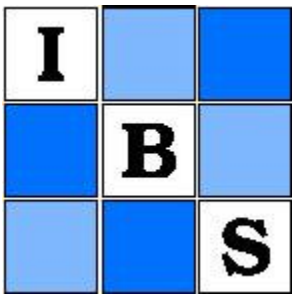

BIOMETRICS *29*, 637–648
December 1973

MAXIMUM LIKELIHOOD ESTIMATION FOR THE BETA-BINOMIAL DISTRIBUTION AND AN APPLICATION TO THE HOUSEHOLD DISTRIBUTION OF THE TOTAL NUMBER OF CASES OF A DISEASE

D. A. GRIFFITHS[1]

Department of Biomathematics, Oxford

SUMMARY

In part I, maximum likelihood (ML) estimation for the beta-binomial distribution (BBD) is considered. The BBD can be used as a model for the incidence in households of noninfectious disease. Typically households in which there are no cases of disease will not be included in the data. It is then necessary to fit a truncated BBD.

Alternative modes of infection are discussed in part II. These give rise to a variety of models for the household distribution of the number of cases of a disease. The BBD is fitted to some data on the common cold and influenza. Other models have been fitted by previous authors to the same data. Independent epidemiological evidence would be necessary for choosing among these models.

I. FITTING THE BETA-BINOMIAL DISTRIBUTION

1. *The Model*

If, in the binomial distribution

$$\phi(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

the probability p of a "success" is allowed to vary between sets of trials with mixing distribution

$$dF(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}\, dp \qquad 0 < p < 1, \quad \alpha, \beta > 0 \tag{1}$$

the resulting distribution has probability function

$$\begin{aligned} P(x) &= \frac{\binom{n}{x} B(\alpha+x, n+\beta-x)}{B(\alpha, \beta)} \\ &= \binom{n}{x} \frac{(\alpha+x-1)(\alpha+x-2) \cdots \alpha(\beta+n-x-1)(\beta+n-x-2) \cdots \beta}{(\alpha+\beta+n-1)(\alpha+\beta+n-2) \cdots (\alpha+\beta)} \\ & \qquad\qquad x = 0, 1, 2, \cdots n. \end{aligned} \tag{2}$$

[1] Present address: C.S.I.R.O. Division of Mathematical Statistics, Sydney, N.S.W., Australia.

This is the beta-binomial distribution (BBD). Its limiting form is the negative binomial (NBD) analogous to the Poisson limit to the binomial. Like the NBD, the BBD can arise in a number of ways (see Moran [1968]) and is sometimes known as the Polya or negative hypergeometric distribution. The mean and variance of the BBD are

$$\mu = n\pi$$
$$\sigma^2 = n\pi\chi(1 + n\theta)/(1 + \theta) \tag{3}$$

where

$$\pi = \frac{\alpha}{\alpha + \beta}, \chi = 1 - \pi, \quad \text{and} \quad \theta = \frac{1}{\alpha + \beta}. \tag{4}$$

Thus $\pi, \theta > 0$. Here π is the mean value of the binomial parameter p and θ is a measure of the variation in p. This reparametrization is convenient especially in cases where a pure binomial model ($\theta = 0$) is thought to be appropriate, and hence both β and α are infinite, which could cause difficulties in estimation and hypothesis testing (see Bailey [1957], section 6.21).

2. *Estimation*

The parameters of the BBD can be estimated by moments. If the observed mean and variance are respectively m and s^2 then the moment estimates are

$$\pi_m = \frac{m}{n} \text{ (and } \chi_m = 1 - \pi_m)$$
$$\theta_m = \frac{s^2 - n\pi_m\chi_m}{n^2\pi_m\chi_m - s^2}. \tag{5}$$

Chatfield and Goodhart [1970] have suggested using the method of mean and zeros to fit the BBD. They proposed a simple iterative technique for calculation of the estimates under this method. The moment estimates are used as starting points and the estimate for θ is adjusted in order to equalize the observed and expected frequencies in the zero class. Shenton [1950] has investigated the efficiency of the method of moments and the results of Anscombe [1950] indicate that for reverse J-shaped distributions (large zero frequency) the method of mean and zeros is more efficient. It should be noted that this comment holds for J-shaped distributions (high frequency in the class $x = n$). Here the method of "mean and nth frequency" for the distribution of the number of "successes" is equivalent to the method of mean and zeros for the distribution of the number of "failures".

Now consider the more efficient method of ML estimation.

Although Haldane [1942] pointed out that it was quite unnecessary to state the ML equations for the NBD in terms of digamma functions, much of the more recent literature has ignored this fact and either used other methods of estimation (e.g. Ehrenberg [1959]) or even laboriously tabulated

functions related to the digamma function in order to 'facilitate' ML estimation (Sichel [1951]).

Precisely the same situation exists with the BBD distribution. Various authors (Ishii and Hayakawa [1960], Chatfield and Goodhart [1970]) have used alternative estimation methods. Skellam [1948] fitted the distribution by maximum likelihood using the digamma function. As with the NBD, the digamma function can be avoided.

Define f_x $(x = 0, 1, 2, \cdots, n)$ as the observed frequencies. Then the likelihood is

$$L(\alpha, \beta) = \prod_{x=0}^{n} [P(x)]^{f_x}.$$

Let $S_i = \sum_{x=0}^{i} f_x$ $(i = 0, 1, 2, \cdots, n)$ so that $S_n = N$, the sample size. Thus, in terms of the parametrization (4), the log likelihood is given by

$$\ln L = c - S_n \sum_{i=1}^{n-1} \ln (1 + i\theta) + \sum_{i=0}^{n-1} \{(S_n - S_i) \ln (\pi + i\theta) + S_{n-1-i} \ln (\chi + i\theta)\}. \quad (6)$$

The ML estimators $\hat{\pi}$, $\hat{\theta}$ are solutions of

$$0 = \frac{\partial \ln L}{\partial \pi}\bigg|_{\hat{\pi},\hat{\theta}} = \sum_{i=0}^{n-1} \left\{\frac{S_n - S_i}{\pi + i\theta} - \frac{S_{n-1-i}}{\chi + i\theta}\right\}$$

$$0 = \frac{\partial \ln L}{\partial \theta}\bigg|_{\hat{\pi},\hat{\theta}} = \sum_{i=1}^{n-1} i\left\{\frac{S_n - S_i}{\pi + i\theta} + \frac{S_{n-1-i}}{\chi + i\theta} - \frac{S_n}{1 + i\theta}\right\}. \quad (7)$$

These may be solved iteratively using the Newton–Raphson method. A convenient starting point is provided by the moment estimates. The second derivatives, and hence the sample information matrix, are readily obtained.

3. *The Truncated BBD*

In applications to household cases of disease (see section 4) the zero class will often not be sampled and data will only be available for those households in which there is at least one case. It is then necessary to fit the zero-truncated BBD. The probability function of the truncated BBD is defined by

$$Q(x) = P(x)/[1 - P(0)] \qquad x = 1, 2, \cdots, n \quad (8)$$

where the $P(x)$ are defined by (2).

For example, if $n = 2$, we have

$$\left.\begin{aligned} Q(1) &= \frac{2\beta}{1 + \alpha + 2\beta} = \frac{2\chi}{1 + \chi + \theta} \\ Q(2) &= \frac{1 + \alpha}{1 + \alpha + 2\beta} = \frac{\pi + \theta}{1 + \chi + \theta} \end{aligned}\right\}. \quad (9)$$

For (8) to represent a probability distribution, it is sufficient that $\theta > 0$, $\pi > -\theta$. Thus, if we ignore the underlying model the restriction that $\pi > 0$ may be relaxed for the distribution defined in (8).

It is quite straight forward to write down from (8) the likelihood and hence solve for $\hat{\pi}$, $\hat{\theta}$ the equations

$$\left.\frac{\partial \ln L}{\partial \pi}\right|_{\hat{\pi},\hat{\theta}} = \left.\frac{\partial \ln L}{\alpha\theta}\right|_{\hat{\pi},\hat{\theta}} = 0.$$

Unconstrained ML estimation could, however, lead to negative estimates for π. Indeed, such is the case with, for example, Lidwell and Somerville's lumped data (Table 1) for which the unconstrained ML parameter estimates are $\hat{\pi} = -0.59$, $\hat{\theta} = 2.53$. For truncated data there is no obvious convenient starting point for ML iteration. One possibility is to use the recurrence relationship

$$R(x) = Q(x+1) - \frac{(\pi + \theta x)(n - x)}{[\chi + \theta(n - x - 1)](x + 1)} Q(x) = 0, \qquad x > 0. \quad (10)$$

Choosing two suitable values of x, (12)—with observed frequencies f_x replacing the probabilities $Q(x)$—yields a pair of simultaneous linear equations for π, θ the solution to which we use as starting point for ML iteration. McGilchrist [1969] has studied the properties of classes of estimators based on recurrence relationships. For the one parameter (τ) case the estimate $\tilde{\tau}$ is the solution of

$$\sum_{x>0} W(x)R(x) = 0. \quad (11)$$

The weights $W(x)$ can be chosen arbitrarily, but McGilchrist proposed a suitable criterion for the choice of optimum weights.

Since a negative estimate of π would be meaningless in terms of the underlying BBD model, the constraint that π is nonnegative is retained here. The above approach to ML estimation for the truncated BBD ignores the problem of constraining π to be nonnegative. An alternative method of estimation, applicable to any truncated discrete distribution, was proposed by McKendrick [1929] although it appears to have been little used in practice (see Irwin [1963]). The method consists in essence of first "guesstimating" the unobserved zero frequency f_0 (or equivalently $N = \sum_{i=0}^{n} f_i$). Then, using this assumed value, the untruncated BBD is estimated by maximum likelihood as outlined in section 2. To obtain the ML estimates for the truncated distribution we iterate until $\hat{P}(0) \simeq f_0/N$ choosing successive values of f_0 to be larger or smaller according to whether $\hat{P}(0)$ is greater than or less than f_0/N. An initial value for f_0 may be found from the observed factorial moments (see Irwin [1963]) although trial and error is easier and, sensibly used, should work well in practice. Hartley [1958] has shown that if a solution is obtained in this way it will be the ML solution. Further, since the parameters of the nontruncated BBD must be nonnegative, this method of estimation effectively solves the problem of constraining $\pi \geq 0$.

The convergence of the iteration process is readily accelerated, as described in section 4 of Hartley [1958]. It may happen that $f_0 \to \infty$ and $\pi \to 0$, corresponding to a situation in which the proportion in the zero class approaches unity. This may not be far from the truth for household cases of a disease which is not common. It may be possible to place an upper limit, m, on f_0 (and hence, effectively, a constraint $\pi \geq \epsilon > 0$) based on knowledge of, say, the number of households in the community from which the data was drawn. This should not make much difference, in practice, to the fit of the BBD to data although, of course, parameter estimates may depend upon the value of m.

An alternative approach to that of fitting the truncated BBD to household disease data would be to fit the untruncated BBD to the "number of additional cases in the household" (that is, omitting the first case). This was not thought to be appropriate in the present study.

4. *Direct Maximization of the Likelihood Function*

As an alternative to the traditional approach to ML estimation for the BBD, a standard function maximization routine such as that of Nelder and Mead [1965] could be used to maximize the log likelihood directly. The moment estimates would provide convenient starting values.

Nelder and Mead's method, incorporating the constraint $\pi > 0$, is especially useful in maximizing the likelihood

$$L = \prod_{x=1}^{n} [Q(x)]^{f_x}$$

of the truncated BBD. This avoids iteration on f_0 as required by the method of McKendrick. Using a "guesstimate" of the unobserved f_0 , the resulting moment estimates could again be used as starting values in the iterative process.

The convergence of this method was quite rapid with the data of part II, especially with good starting values (that is, a good "guesstimate" of f_0). For example, with Lidwell and Somerville's lumped data a "guesstimate" of $f_0 = 1000$ leads to moment estimates $\pi_m = 0.06$, $\theta_m = 1.06$ and convergence to an accuracy of 10^{-2} after 33 iterations. Even with poor starting values, e.g. $\pi_m = 0.5$, $\theta_m = 0.1$ (corresponding to $f_0 = 0$) convergence still took only 46 iterations.

II. INFECTIOUSNESS AND THE HOUSEHOLD DISTRIBUTION OF DISEASE CASES—AN APPLICATION OF THE BBD

1. *Detection of Infectiousness of a Disease*

There is considerable variation in the literature on the precise meaning attached to such terms as "communicable," "contagious," and "infectious". We will take them to be synonymous and define a disease as infectious if it is capable of transmission from case to case (that is, infection of a sus-

ceptible by an infective individual *during the course of an epidemic outbreak* such that a result of a chain of such case to case transmission is the spread of infection through the population—giving rise to an epidemic).

If a disease is highly infectious its epidemiological pattern may be such as to allow immediate recognition of infectiousness, especially if the disease exhibits a relatively constant serial interval. Alternatively, and for less infectious diseases, careful epidemiological surveillance may lead to the detection of infectiousness.

If no firm conclusions can be reached in this way, some statistical test of significance on suitably collected data may be useful in making inferences about infectiousness. A number of statistical techniques have been applied to the detection of clustering of disease cases in space and time (see, for example, Barton *et al.* [1965], Knox [1964]). Such clustering is circumstantial evidence in favor of contagion, although other interpretations are possible. These methods are appropriate to fairly rare diseases, leukaemia for example.

For more common diseases the household may provide a useful unit for observation.

2. *The Household Distribution of Cases*

With household disease data the times of occurrence (and, if the diseases are infectious, the links in the "chain of infection") will quite commonly be unobservable. Any statistical study must then be based on the observed household distribution of cases.

For those infectious diseases such as measles, which are characterized by a virtually constant latent period followed by a relatively short period of infectiousness, a chain-binomial model is appropriate. Bailey [1957] has considered various chain-binomial models including the case where the chain-binomial parameter (as with the ordinary binomial parameter in the BBD) varies among households according to the β distribution.

If the period of infectivity is not confined to a short period of time (in chain-binomial models this period is reduced to a single point), then a continuous-infection model (see Bailey [1957]) is more appropriate. Such a model might apply to diseases such as scarlet fever or diphtheria.

It is assumed in the above that, after the introduction of one or more primary cases into the household, infection is propagated from case to case within the household and the possibility of further infection entering from outside is excluded.

Suppose, however, that a source which is external to the household or which is internal but environmental and independent of the existence of infectious individuals is partly (perhaps wholly) responsible for the continuation of the household "epidemic". McKendrick [1926] first investigated this problem and was able to estimate the ratio of the probability of internal infection to the probability of external infection from the observed total number of cases, the total number of houses in the community and the total number of infected houses. Applying this to four epidemics of bubonic plague in a village, he concluded that the ratio was about 200 : 1.

The chain-binomial model of Sugiyama [1961] allows for infection from outside the household, which may involve person-to-person contact as well as case-to-case infection within the household. The relative importance of the two modes of infection could be assessed by estimating parameters of these models from suitable observed household data. Sugiyama's model is somewhat unrealistic in that it assumes that infection may only enter the household at discrete time points corresponding to the serial interval. This restriction may not seriously affect the distribution of the total number of cases. Although Sugiyama appears to have successfully separated into the constituent chains data on the household distribution of influenza cases, it would seem better, in general, to base any statistical analysis on the distribution of the total number of cases. Given that there may be some difficulty in distinguishing the links in the chain of infection for a "closed" household, it does seem rather ambitious to hope that these links can be distinguished under Sugiyama's model. The analogous continuous-infection model, or some other model involving the two modes of infection, may be worth further investigation.

For many respiratory infections such a model may be appropriate. Though it is known (Hope–Simpson [1958], Andrewes [1965]) that colds can be caught, the importance of case-to-case transmission is not known, and may vary among the many different viruses which may cause acute afebrile upper respiratory infection (i.e. the common cold). Epidemiological evidence, including the observed simultaneous occurrence of waves of infection in different regions, has led to the suggestion that person-to-person spread seems inconceivable as a full explanation of the incidence of the common cold and influenza (Hope–Simpson [1958; 1968]). An interesting explanatory hypothesis has been suggested (see Hope–Simpson [1968], Andrewes [1967]). Suppose that at the halt of an outbreak a number of individuals have been seeded with virus but that environmental circumstances do not permit virus multiplication, so that the individuals will not be infectious nor will they show any clinical symptoms. After some time, environmental circumstances will change and some trigger (perhaps meteorological) will permit the arrested process to resume. This, then, is the beginning (perhaps the whole?) of the next outbreak. That weather may act as a trigger which activates a latent infection in the individual has been shown, for swine influenza virus, by Shope [1955].

In accord with the definition in section 1 of an infectious disease, we categorize those diseases involving activated latent infection as noninfectious. For a disease which is not infectious, the distribution of the total number of cases arising in households of a given size will be binomial if cases occur at random. Allowing for variation between households we get the BBD. If the zero class is unobserved and we only have data on those households in which there is at least one case, the truncated BBD will apply. Irwin [1954] derived the BBD on the assumption of serial dependence in the probabilities of selection and applied the model to measles incidence in households. In view of the epidemiology of measles, which is a highly infectious

disease, it would seem to be an inappropriate application although a good fit was obtained.

3. *Application to Household Disease Data*

Spicer (discussion on Bailey [1955]) reported a good fit of the BBD to data on Sonne dysentry. We consider here some data on respiratory infections.

Lidwell and Somerville [1951] collected data on the incidence of the common cold. To remove a possible source of heterogeneity, households were classified according to their type. For households of size 4, which comprise the bulk of the data, the possible households were classified according to whether they contained

(i) adults only
(ii) adults and schoolchildren
(iii) adults and infants
(iv) adults, schoolchildren and infants.

The observed household distributions of cases for the Lidwell–Somerville data are presented in Table 1. To test for heterogeneity, a χ^2-test was performed on the 4×4 array of Table 1. The χ^2 value of 11.9 on 9 D.F. ($P = 0.22$) indicates that there is no significant departure from homogeneity. It seems reasonable therefore to lump the data.

The truncated BBD was fitted by maximum likelihood, using the method of McKendrick [1926], to each of the four sets of data in Table 1, as well as the lumped data. The resulting fits, as well as that of a truncated "pure" binomial distribution, are compared in Table 2. Because of the small expectations the "χ^2" values are only very roughly distributed as χ^2. However, it is quite clear that the truncated binomial model is inappropriate. By contrast, none of the χ^2 values for the fitted BBD distributions is significant at the $2\frac{1}{2}\%$ level. It is obvious that the BBD fits the data quite well.

TABLE 1

THE OBSERVED DISTRIBUTION OF TOTAL NUMBER OF CASES OF COMMON COLD IN HOUSEHOLDS OF SIZE 4 (LIDWELL–SOMERVILLE)

Household type	Number of cases			
	1	2	3	4
(i)	15	5	2	2
(ii)	12	6	7	6
(iii)	10	9	2	7
(iv)	26	15	3	9
lumped data (total)	63	35	14	24

TABLE 2

COMPARISON OF FITTED DISTRIBUTION (TRUNCATED BBD AND TRUNCATED BINOMIAL) AND OBSERVED DATA OF LIDWELL AND SOMERVILLE

Number of cases		1	2	3	4	$\hat{\pi}$	$\hat{\theta}$	χ^2	d.f.
Type (i) Households	Observed	15	5	2	2				
	Truncated BBD	13.9	5.8	2.9	1.4	0	0.62	0.8	1
	Truncated Binomial	14.0	7.8	1.9	0.2	0.27	-	2.9	1
Type (ii) Households	Observed	12	6	7	6				
	Truncated BBD	11.6	7.2	5.9	6.3	0.15	1.16	0.4	1
	Truncated Binomial	6.9	12.0	9.3	2.7	0.54	-	11.4	2
Type (iii) Households	Observed	10	9	2	7				
	Truncated BBD	10.6	6.1	5.1	6.3	0	1.65	3.4	1
	Truncated Binomial	6.3	10.9	8.4	2.4	0.53	-	16.1	2
Type (iv) Households	Observed	26	15	3	9				
	Truncated BBD	24.8	12.6	8.7	6.9	0	1.06	4.9	1
	Truncated Binomial	24.8	20.0	7.2	1.0	0.35	-	68	2
Lumped Data	Observed	63	35	14	24				
	Truncated BBD	60.4	31.9	23.1	20.6	0	1.19	4.6	1
	Truncated Binomial	59.4	52.7	20.8	3.1	0.37	-	151	2

TABLE 3

OBSERVED AND FITTED DISTRIBUTIONS—DATA OF HEASMAN AND REID

Number of cases	Uncrowded		Crowded		Overcrowded		Total	
	Obs.	Fitted	Obs.	Fitted	Obs.	Fitted	Obs.	Fitted
1	156	156.4	155	142.7	112	104.3	423	402.5
2	55	53.2	41	55.1	35	41.6	131	150.5
3	19	21.7	24	26.3	17	20.7	60	69.2
4	10	8.4	15	12.4	11	10.3	36	31.0
5	2	2.3	6	4.5	6	4.1	14	10.5
χ^2(2 d.f.)	0.8		5.9		3.2		6.7	
$\hat{\theta}$	0.34		0.46		0.49		0.43	
$\hat{\pi}$	0		0		0		0	

Heasman and Reid [1961] have examined data on family epidemics of the common cold. Their data was subdivided according to the degree of crowding and is presented (for families of size 5) in Table 3. Again, a χ^2-test showed no significant departure from homogeneity. The truncated BBD also fits this data (lumped and unlumped) reasonably well.

There seems to be very little further suitable data in the literature. Influenza data of Hope–Simpson and Sutherland [1954] and Sugiyama [1961] are presented in Table 4. In both cases the BBD fits the data quite well.

Heasman and Reid obtained good fits with both chain-binomial and continuous-infection models. Lidwell and Somerville, and also Hope–Simpson and Sutherland, also obtained good fits to their data with chain-binomial distributions, as did Sugiyama with his modified chain-binomial model. One can only concur with the comment of Spicer (discussion on Bailey [1955]):

> "I do not know how you choose between these models; ... you can explain the data on widely differing mathematical hypotheses, and there is very little else you can do"

without further and *independent* epidemiological evidence.

TABLE 4

OBSERVED AND FITTED DISTRIBUTIONS—DATA OF HOPE–SIMPSON AND SUTHERLAND, AND SUGIYAMA

Data Source	Number of Cases						
	0	1	2	3	4	5	
Hope-Simpson and Sutherland							
(Households of size 4)	-	3	3	6	12	-	Observed
$\hat{\pi} = 0.72$, $\hat{\theta} = 0.58$	-	2.7	3.7	5.4	12.1	-	BBD
$\chi^2 = 0.2$ (1 d.f.)							
(Households of size 5)	-	1	2	2	4	9	Observed
$\hat{\pi} = 0.77$, $\hat{\theta} = 0.48$	-	1.1	1.7	2.4	3.8	9.0	BBD
$\chi^2 = 0.2$ (2 d.f.)							
Sugiyama							
(Households of size 3 including zero class)	29	9	2	2	-	-	Observed
$\hat{\pi} = 0.15$, $\hat{\theta} = 0.43$	29.2	7.8	3.6	1.4	-	-	BBD
$\chi^2 = 1.1$ (1 d.f.)							

ACKNOWLEDGMENTS

I should like to thank Professor M. S. Bartlett who supervised this research. I am grateful to Professor J. B. Douglas and Dr. C. A. McGilchrist for helpful discussion and to Dr. R. E. Hope-Simpson for his advice on the epidemiology of respiratory infections. The critical comments of the referees have led to considerable revision and improvement of the original manuscript.

ESTIMATION DU MAXIMUM DE VRAISEMBLANCE POUR LA DISTRIBUTION BETA-BINOMIALE ET APPLICATION A LA REPARTITION DANS LES MENAGES DU NOMBRE DE CAS D'UNE MALADIE

RESUME

Dans la partie I, on étudie l'estimation du maximum de vraisemblance pour la distribution beta-binomiale (BBD). La BBD peut être utilisée comme modèle pour les nombres de cas d'une maladie non-infectieuse dans les ménages. L'ajustement à une BBD tronquée est rendue nécessaire par l'absence du décompte des ménages sans un seul cas de la maladie.

Différents modes d'infection sont discutés en partie II. Ils conduisent à un choix de modèles pour la distribution du nombre de cas d'une maladie dans les ménages. On a ajusté la BBD à certaines données concernant le rhume commun et la grippe. D'autres modèles ont été ajustés antérieurement aux mêmes données. Des preuves indépendantes de nature épidémiologique sont nécessaires pour choisir entre ces modèles.

REFERENCES

Andrewes, C. H. [1965]. *The Common Cold.* Weidenfeld and Nicholson, London.

Andrewes, C. H. [1967]. *The Natural History of Viruses.* Weidenfeld and Nicholson, London.

Anscombe, F. J. [1950]. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika 37*, 358–82.

Bailey, N. T. J. [1955]. Some problems in the statistical analysis of epidemic data. *J. R. Statist. Soc. B17*, 35–68.

Bailey, N. T. J. [1957]. *The Mathematical Theory of Epidemics.* Griffin, London.

Barton, D. E., David, F. N., and Merrington, M. [1965]. A criterion for testing contagion in time and space. *Ann. Hum. Genet. 29*, 97–102.

Chatfield, C. and Goodhart, G. J. [1970]. The beta-binomial model for consumer purchasing behaviour. *Appl. Statist. 19*, 240–50.

Ehrenberg, A. S. C. [1959]. The pattern of consumer purchases. *Appl. Statist. 8*, 26–41.

Haldane, J. B. S. [1942]. The fitting of binomial distributions. *Annals of Eugenics 11*, 179–81.

Hartley, H. O. [1958]. Maximum likelihood estimates from incomplete data. *Biometrics 14*, 174–94.

Heasman, M. A. and Reid, D. D. [1961]. Theory and observation in family epidemics of the common cold. *Br. J. Prev. Soc. Med. 15*, 12–6.

Hope–Simpson, R. E. [1958]. Common upper respiratory diseases. *Roy. Soc. Hlth. J. 78*, 593–9.

Hope–Simpson, R. E. [1968]. The natural history of some common viruses. Paper read at B.M.A. Conference, Cheltenham, 25.10.68.

Hope–Simpson, R. E. and Sutherland, I. [1954]. Does influenza spread within the household? *The Lancet*, 721–6.

Irwin, J. O. [1954]. A distribution arising in the study of infectious diseases. *Biometrika 41*, 266–8.

Irwin, J. O. [1963]. The place of mathematics in medical and biological statistics. *J. R. Statist. Soc. A126*, 1–44.

Ishii, G. and Hayakawa, R. [1960]. On the compound binomial distribution. *Ann. Inst. of Statist. Math. 12*, 69–80.

Knox, E. G. [1964]. The detection of space-time interactions. *Appl. Statist. 13*, 25–9.

Lidwell, O. M. and Somerville, T. [1951]. Observations on the incidence and distribution of the common cold in a rural community during 1948 and 1949. *J. Hyg. Camb. 49*, 365–81.

McGilchrist, C. A. [1969]. Discrete distribution estimators from the recurrence equation for probabilities. *J. Amer. Statist. Ass. 64*, 602–9.

McKendrick, A. G. [1926]. Applications of mathematics to medical problems. *Proc. Edin. Math. Soc. 44*, 98–130.

Moran, P. A. P. [1968]. *An Introduction to Probability Theory.* Oxford University Press.

Nelder, J. A. and Mead, R. [1965]. A simplex method for function minimisation. *The Computer Journal 7*, 308–13.

Shenton, L. R. [1950]. Maximum likelihood and the efficiency of the method of moments. *Biometrika 37*, 111–6.

Shope, R. E. [1955]. The swine influenza lungworm as a reservoir and intermediate host for swine influenza virus. V. Provocation of swine influenza by exposure of prepared swine to adverse weather. *J. Exp. Med. 102*, 567–72.

Sichel, H. S. [1951]. The estimation of the parameters of a negative binomial distribution with special reference to psychological data. *Psychometrika 16*, 107–27.

Skellam, J. G. [1948]. A probability distribution derived from the binomial distribution by regarding the probability of a success as variable between the sets of trials. *J. R. Statist. Soc. B10*, 257–61.

Sugiyama, H. [1961]. Some statistical methodologies for epidemiological research of medical sciences. *Bull. Int. Statist. Inst. 38(3)*, 137–51.

Received June 1971, *Revised August* 1972

Key Words: Beta-binomial distribution; Maximum likelihood estimation; Truncated BBD; Chain binomial distribution; Infectious diseases—household data; Infectiousness of influenza and the common cold.