

In what follows, the data are analysed in accordance with the theory developed for bivariate models as in Chapter 2. Use is made of the ascending diagonal arrays and the third marginal distribution. For the ascending diagonal arrays we fit the BBD and, from the estimation of the parameters α_1 and α_2 , we observe that it is not possible to employ the same distribution along all the arrays. From Figure 9.22 one can see that $\hat{\alpha}_1$ and $\hat{\alpha}_2$ increase with S and that the same function is suitable for both parameters, subject to a constant. By calculating weighted regression curves to α_1 and α_2 , we obtain the following functions:

$$\hat{\alpha}_1(S) = 0,4290088 S + 2,44629 \quad \text{and}$$

$$\hat{\alpha}_2(S) = 0,4290088 S + 1,967816.$$

For the third marginal we fit the zero-augmented NBD using the method of moments. The estimates obtained are:

$$\hat{\tau} = 1,838088, \quad \hat{\theta} = 0,871953 \quad \text{and} \quad \hat{q} = 7,145728 \times 10^{-3}.$$

As the sample size is extremely large, one should not be surprised about the poor performance of the χ^2 test, as shown in Table 9.13. From the graphs of Figures 9.23, 9.24 and 9.25 one can see that excellent fits of the empirical data have been obtained.

Table 9.13 The χ^2 results obtained from Table 9.12

	χ^2	d.f.	$P(\chi^2)$
x_1 marginal	23,63	15	0,1
x_2 marginal	128,65	15	0
Third marginal	68,60	26	$1,1 \times 10^{-5}$

The graphical displays of the x_1 marginal, the x_2 marginal and the third marginal are presented in Figures 9.23, 9.24 and 9.25 respectively.

Figure 9.26 shows the empirical and the theoretical regression curve $E(x_2|x_1)$.

The correlation coefficient between the variables x_1 and x_2 derived from the theoretical fit is $\rho = 0,631$ and the one obtained from the empirical data is 0,635.

Figure 9.22 Estimates of α_1 and α_2 as a function of S ,
data of Cane and Burrell (1982)

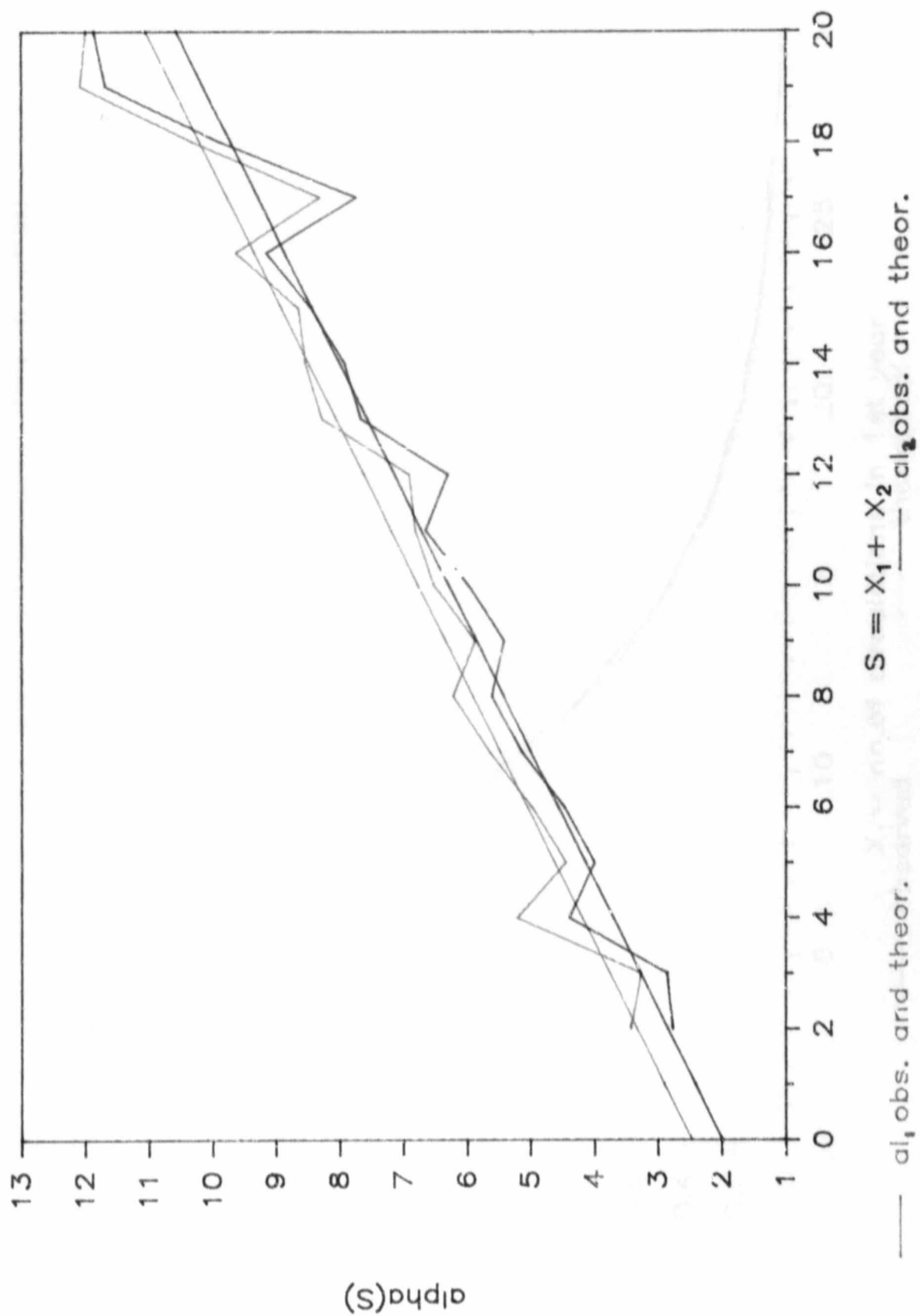


Figure 9.23 The x_1 marginal, data of Cane and Burrell (1982)

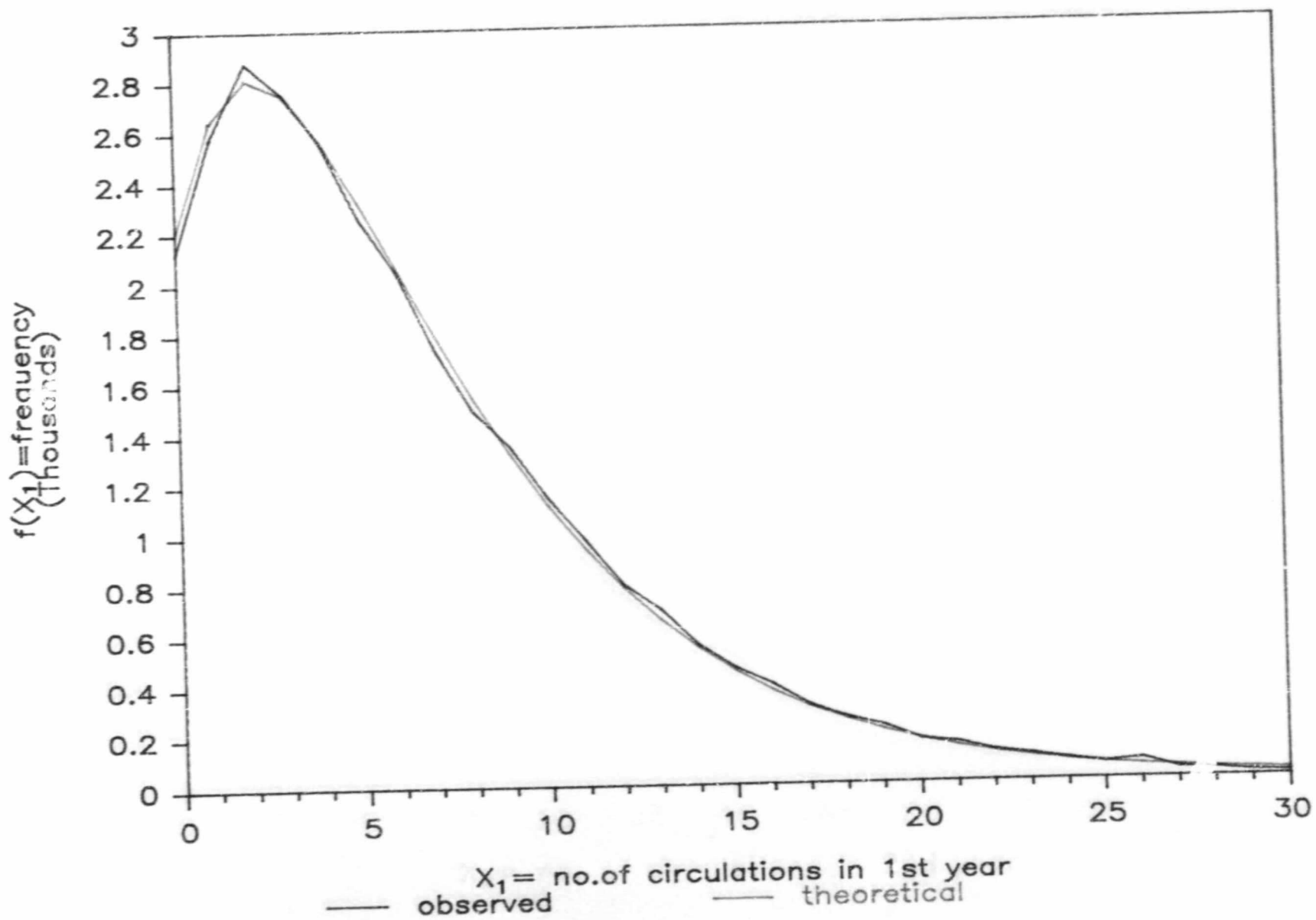


Figure 9.24 The x_2 marginal, data of Cane and Purrell (1982)

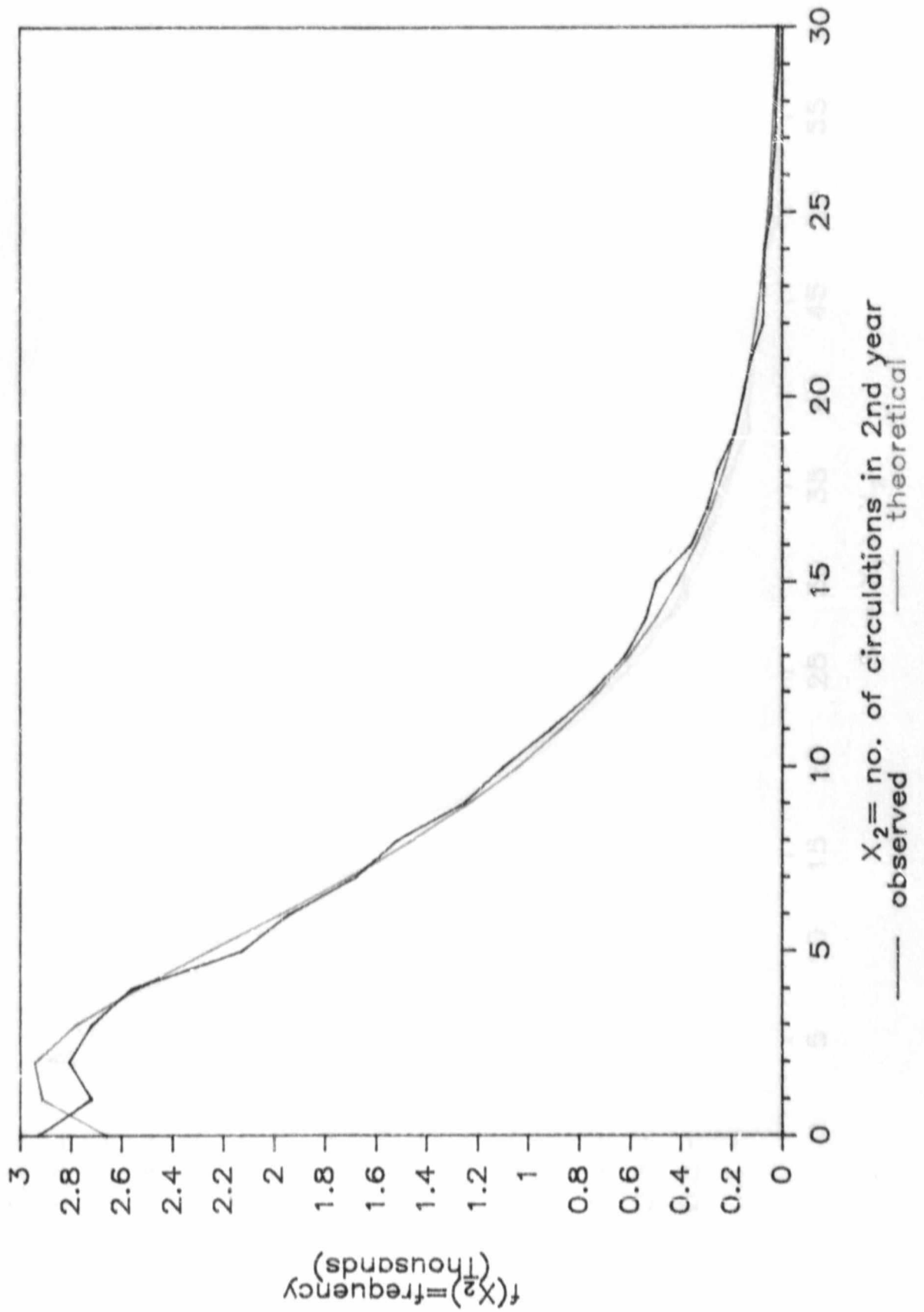


Figure 9.25 The third marginal, data of Cane and Burrell (1982)

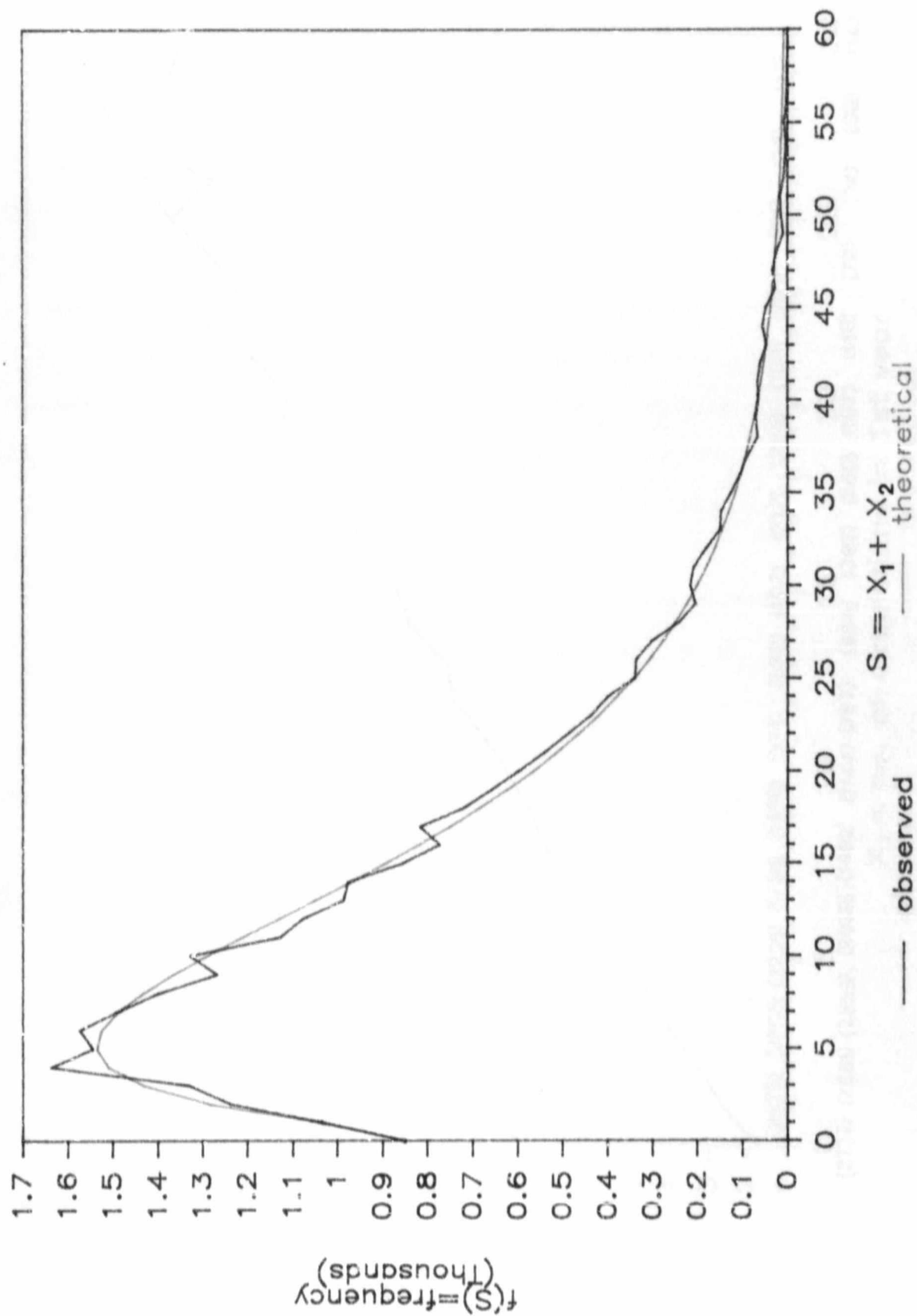
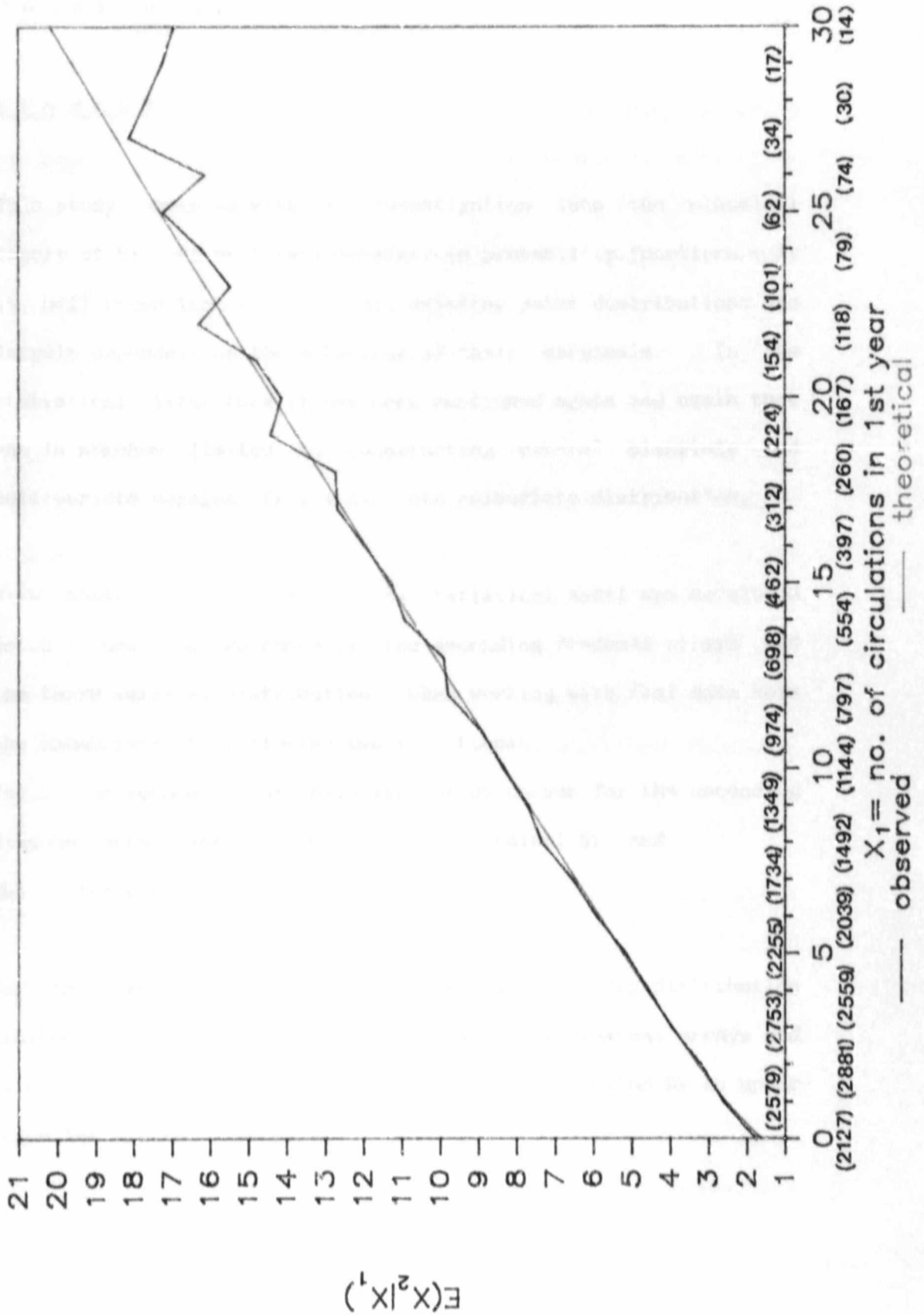


Figure 9.26 The regression curve $E(x_2|x_1)$,
data of Camo and Burrell (1982)



CHAPTER 10

SUMMARY

This study commenced with an investigation into the classical theory of bi- and multivariate discrete probability functions. It is well known that virtually all existing joint distributions are largely dependent on the behaviour of their marginals. In the statistical literature it has been mentioned again and again that one is somehow limited in constructing general bivariate and multivariate versions from a discrete univariate distribution.

With these points in mind a new statistical model was developed which is based on two concepts: the ascending diagonal arrays and the third marginal distribution. When working with real data sets one encounters the following two situations:

- (a) the parameters of the distribution chosen for the ascending diagonal arrays are independent of the terminal S_i and
- (b) the parameters are dependent on S_i .

For the bivariate case the model makes use of any distribution limited by an upper terminal for the ascending diagonal arrays and it employs any distribution which is or is not limited by an upper value for the third marginal.

When the distribution applied for the third marginal is limited by an upper terminal, the joint distribution is presently bounded by the main ascending diagonal array, although future research may overcome this limitation.

As a result of the process of combining the third marginal with the ascending diagonal arrays, a variety of new bivariate families of distributions was created. In view of the fact that these families are built with the help of the ascending diagonal arrays, a greater flexibility is introduced which is lacking in the classical discrete bivariate distributions.

In order to illustrate the applications of the general model, some families of distributions were considered for the ascending diagonal arrays and for the third marginal.

The new model allows for the calculation of moments for the marginals x_1 and x_2 and also for the correlations between the variables, without explicitly deriving the probability functions for the marginals.

In certain cases it is possible to work out general expressions for the regression curves. Obviously, the derivation of analytic formulae for the regression curves necessitates the knowledge of the marginal distributions. The NBD-BBD model was presented and for certain functional relations between the parameters a few families of regressions were developed. The use of the Beta-Binomial distribution (BBD), the Generalized Positive

Hypergeometric distribution (GPHD) and the Binomial distribution for the ascending diagonal arrays allows for linear, convex and concave regression curves.

The general trivariate case was constructed on the same principle as the bivariate model. A new array, called the total of variables, was defined. Similar to the bivariate model, this concept enables the use of any distribution for the ascending diagonal arrays, the third marginal and the total of variables.

The multivariate case was further derived by a reduction to a number of trivariate models. As an example for the multivariate case we took the Dirichlet-Multinomial distribution. This probability law is known for its low correlations between the variables. By applying any distribution on the total of variables, it is possible to obtain a much increased range of correlations. Specifically we took the NBD as the distribution of the total of variables. By using different values for the parameters γ , θ and $\alpha_0, \dots, \alpha_k$ we demonstrated the wide range of correlations which may be obtained.

The general properties of Poisson mixtures used for the third marginal were investigated. Special attention was given to the behaviour of their moments as functions of the exposure period t and of the moments of the mixing distribution.

Some statistical indices of major importance were studied, such as

Hypergeometric distribution (GPHD) and the Binomial distribution for the ascending diagonal arrays allows for linear, convex and concave regression curves.

The general trivariate case was constructed on the same principle as the bivariate model. A new array, called the total of variables, was defined. Similar to the bivariate model, this concept enables the use of any distribution for the ascending diagonal arrays, the third marginal and the total of variables.

The multivariate case was further derived by a reduction to a number of trivariate models. As an example for the multivariate case we took the Dirichlet-Multinomial distribution. This probability law is known for its low correlations between the variables. By applying any distribution on the total of variables, it is possible to obtain a much increased range of correlations. Specifically we took the NBD as the distribution of the total of variables. By using different values for the parameters γ , θ and $\alpha_0, \dots, \alpha_k$ we demonstrated the wide range of correlations which may be obtained.

The general properties of Poisson mixtures used for the third marginal were investigated. Special attention was given to the behaviour of their moments as functions of the exposure period t and of the moments of the mixing distribution.

Some statistical indices of major importance were studied, such as

the coefficients of variation, dispersion, skewness and kurtosis and the Yule characteristic. They were expressed in terms of the moments of the mixing distribution and the exposure period t . For $t \rightarrow \infty$ the coefficients of skewness and kurtosis of the compound Poisson distribution tend to the respective coefficients of the mixing distribution. This was a rather surprising result, as one might have expected a tendency to normality.

Some of the results obtained for the general Compound Poisson distributions were applied to the Generalized Inverse Gaussian-Poisson, the Inverse Gaussian-Poisson and the Negative Binomial distributions. These distributions have been investigated in detail and their behaviour as a function of t was presented.

In addition, the bivariate, trivariate and multivariate models were considered in terms of the exposure period t as well as their limiting process. When we analysed the bivariate, trivariate and multivariate models as a function of time we found some similarities between the correlation coefficients at $t = 1$ and at $t \rightarrow \infty$.

In order to illustrate the new theory developed herein, some specific models have been applied to several data sets. A few estimation procedures were presented, some of them well known in the statistical literature. An attempt was made to derive the likelihood functions for some of the new models, including the derivation of the maximum likelihood equations. These can not

always be obtained in a closed form and their solution requires some numerical methods.

By calculating the expectation of the negative second derivatives of the likelihood functions we obtained the expected Fisher information matrix.

The application of the above methods to four different data sets gave satisfactory fits and in some cases the results were better than the ones achieved with models which had previously been proposed in the statistical literature.

The data set of Bates and Neyman (1952) from the accident field was fitted with the NBD-BBD model. The model assumed the BBD with the same parameters α_1 and α_2 along all the arrays and the NBD for the third marginal. The goodness of fit was measured in terms of the χ^2 test, graphical fittings, a comparison between the observed and the theoretical correlations and a comparison between the observed and the theoretical regression curves.

The data set for accidents of children used by Mellinger et al (1965) was also fitted with the NBD - BBD model. Agreement with the theory was tested with the same measures as described above. Satisfactory results were obtained.

From the library circulation field we investigated the data set of Chen (1976). We were able to fit two different models, namely the NBD - GPHD and the NBD - Binomial. Using the NBD - GPHD a concave

always be obtained in a closed form and their solution requires some numerical methods.

By calculating the expectation of the negative second derivatives of the likelihood functions we obtained the expected Fisher information matrix.

The application of the above methods to four different data sets gave satisfactory fits and in some cases the results were better than the ones achieved with models which had previously been proposed in the statistical literature.

The data set of Bates and Neyman (1952) from the accident field was fitted with the NBD-BBD model. The model assumed the BBD with the same parameters α_1 and α_2 along all the arrays and the NBD for the third marginal. The goodness of fit was measured in terms of the χ^2 test, graphical fittings, a comparison between the observed and the theoretical correlations and a comparison between the observed and the theoretical regression curves.

The data set for accidents of children used by Mellinger et al (1965) was also fitted with the NBD - BBD model. Agreement with the theory was tested with the same measures as described above. Satisfactory results were obtained.

From the library circulation field we investigated the data set of Chen (1976). We were able to fit two different models, namely the NBD - GPHD and the NBD - Binomial. Using the NBD - GPHD a concave

regression curve was estimated. The NBD - Binomial model gave a slightly better result.

The new model was also applied to data of Cane and Burrell (1982) dealing with library book circulation. Here it was necessary to express the parameters α_1 and α_2 as functions of S . The model chosen was the NBD - BBD. This very large data set is quite unique and it is difficult to find an appropriate model such as to provide a good fit in terms of the χ^2 test.

Further research

The newly developed theory could be followed up with some additional research. A few directions are suggested for further investigation:

(a) Where the distribution for the third marginal was limited by an upper terminal, only bivariate distributions which are bounded by the main ascending diagonal array were obtained. More generalized joint distributions should be developed.

(b) For the case where the parameters for the ascending diagonal arrays can not be regarded as constant along the whole bivariate table, different functions should be considered.

(c) For various joint distributions with certain functional relations between their parameters, new regression curves could be developed.

(d) A further step could be taken in the inference area, related to these joint distributions. More likelihood functions could be derived for different joint distributions and, by using numerical methods such as the Newton-Raphson, the solutions for the maximum likelihood equations may be obtained.

(e) Further research could also be undertaken in the calculation of Fisher's expected information matrix.

REFERENCES

Abramowitz, M. and Stegun, I.A. (1965) *Handbook of Mathematical Functions*, New York: Dover Publications.

Aitchison, J. (1982) The statistical analysis of compositional data, *Journal of the Royal Statistical Society, Ser. B*, 44, No. 2, pp. 139-177.

Anscombe, F.J. (1950) Sampling theory of the negative binomial and logarithmic series distributions, *Biometrika*, 37, pp. 358-382.

Arbous, A.G. and Kerrich, J.E. (1951) Accident statistics and the concept of accident proneness, *Biometrics*, 7, pp. 340-432.

Atkinson, A.C. and Lam Yeh (1982) Inference for Sichel's Compound Poisson distribution, *Journal of the American Statistical Association*, 77, pp. 153-158.

Bartko, J.J. (1961) The negative binomial distribution: a review of properties and applications, *Virginia Journal of Science*, 2, pp. 18-37.

Bates, G.E. and Neyman, J. (1952) Contributions of the theory of accident proneness, I: an optimistic model of the correlation between light and severe accidents, *University of California Publications in Statistics*, 1, pp. 215-253.

Bulmer, M.G. (1974) On fitting the Poisson-Lognormal distribution to species abundance data, *Biometrics*, 30, pp. 101-110.

Burrell, Q.L. and Cane, V.R. (1982) The analysis of library data, *Journal of the Royal Statistical Society, Ser. A*, 145, Part 4, pp. 439-471.

Burrell, Q.L. (1986) *Private communication*.

Chatfield, C., Ehrenberg, A.S.C. and Goodhardt, G.J. (1966) Progress on a simplified model of stationary purchasing behaviour, *Journal of the Royal Statistical Society, Ser. A*, 129, pp. 317-367.

Chen, C.C. (1976) Applications of operations research models to libraries: A case study of use of monographs in the Countway library of medicine, M.I.T. Press, Cambridge, Massachusetts.

Connor, R.J. and Mosimann, J.E. (1969) Concepts of independence for proportions with a generalization of the Dirichlet distribution, *Journal of the American Statistical Association*, 64, pp. 194-206.

Darroch, J.N. and James, I.R. (1974) F-independence and null correlations of continuous, bounded-sum, positive variables. *Journal of the Royal Statistical Society, Ser. B.*, 36, pp. 467-483.

Darwin, J.H. (1960) An ecological distribution akin to Fisher's logarithmic distribution. *Biometrics*, 16(1), pp. 51-60,

Ehrenberg, A.S.C. (1959) The pattern of consumer purchases, *Applied Statistics*, 8, pp. 25-41.

Ehrenberg, A.S.C. (1968) The practical meaning of usefulness of the NBD/LSD theory of repeat buying. *Applied Statistics*, 17, pp. 17-32.

Ehrenberg, A.S.C. (1972) *Repeat-Buying: Theory and Applications*, Amsterdam: North-Holland Publishing Co.

Fabius, J. (1973) Two characterizations of the Dirichlet distribution. *The Annals of Statistics*, Vol. 1, No. 3, pp. 583-587.

Firer, C. (1985) Modelling multibrand consumption. *Omega Int Journal of Management Science*, Vol 13, No. 4, pp. 285-294.

Fisher, R.A. (1941) The negative binomial distribution, *Ann. Eugen.*, London, 11, p.182.

Darroch, J.N. and James, I.R. (1974) F-independence and null correlations of continuous, bounded-sum, positive variables. *Journal of the Royal Statistical Society, Ser. B.*, 36, pp. 467-483.

Darwin, J.H. (1960) An ecological distribution akin to Fisher's logarithmic distribution. *Biometrics*, 16(1), pp. 51-60,

Ehrenberg, A.S.C. (1959) The pattern of consumer purchases, *Applied Statistics*, 8, pp. 26-41.

Ehrenberg, A.S.C. (1968) The practical meaning of usefulness of the NBD/LSD theory of repeat buying. *Applied Statistics*, 17, pp. 17-32.

Ehrenberg, A.S.C. (1972) *Repeat-Buying: Theory and Applications*. Amsterdam: North-Holland Publishing Co.

Fabius, J. (1973) Two characterizations of the Dirichlet distribution. *The Annals of Statistics*, Vol. 1, No. 3, pp. 583-587.

Firer, C. (1985) Modelling multibrand consumption. *Omega Int Journal of Management Science*, Vol 13, No. 4, pp. 285-294.

Fisher, R.A. (1941) The negative binomial distribution, *Ann. Eugen.*, London, 11, p.182.

Fisher, R.A. *Statistical Methods For Research Workers*, 12th ed., London, Section 18, Example 6.

Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population, *Journal of Animal Ecology*, Vol 12, pp. 42-57.

Folks, J.L. and Chikara, R.S. (1978) The Inverse Gaussian distribution and its statistical application - a review (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 40, pp. 263-289.

Gart, J.J. (1958) *Some Problems In Statistical Inference*, Oak Ridge National Laboratory, Oak Ridge, Tenn., Report ORNL-2629.

Goodhardt, G.J., Ehrenberg, A.S.C. and Chatfield, C. (1984) The Dirichlet: A comprehensive model of buying behaviour, *Journal of the Royal Statistical Society, Ser. A*, 147, Part 5.

Grahn, G.L. (1969) The negative binomial distribution model of repeat purchasing loyalty. An empirical investigation, *Journal of Marketing Research*, 6, pp. 72-78.

Greenwood, M. and Yule, G.U. (1920) An enquiry into the nature of frequency distributions representative of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society*, 83, pp. 255-279.

Griffiths, D.A. (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, *Biometrics*, 29, pp. 637-648.

Grundy, P.M. (1951) The expected frequencies in a sample of an animal population in which the abundances of species are lognormally distributed, Part I, *Biometrika*, 38, pp. 427-434.

Gurland, J. (1954) Some applications of the negative binomial and other contagious distributions, *American Journal of Public Health*, 49, pp. 1388-1399.

Haldane, J.B.S. (1941) The fitting of binomial distributions, *Ann. Eugen.*, London, 11, p. 179.

Ishii, G. and Hayakawa, R. (1960) On the compound binomial distribution, *Ann. Inst. of Statistical Mathematics*, 12, pp. 69-80.

James, I.R. and Mosimann J.E. (1980) A new characterization of the Dirichlet distribution through neutrality, *The Annals of Statistics*, 8, pp. 183-189.

James, I.R. (1981) Distributions associated with neutrality properties for random proportions, *Statistical Distributions in Scientific Work*, (Taillie, C., Patil, G.P. and Baldessari, B., editors), Dordrecht, Holland: D. Reidel Publishing Company, Vol. 4, pp. 125-136.

Johnson N.L. (1957) Uniqueness of a result in the theory of accident proneness, *Biometrika*, Vol. 44, pp. 530-531.

Johnson, N.L. and Kotz, S. (1969) *Discrete Distributions*, New York: John Wiley.

Jørgensen, B. (1982) Statistical properties of the Generalized Inverse Gaussian Distribution, *Lecture Notes In Statistics*, 9, New York: Springer-Verlag.

Katz, L. (1965) Unified treatment of a broad class of discrete probability distributions, *Classical and Contagious Discrete Distributions*, Ed. G.P. Patil, Pergamon Press, Oxford, 175.

Kemp, C.D. and Papageorgiou, H. (1982) Bivariate Hermite Distributions, *Sankhya*, 44, Ser. A., Pt. 2, pp. 269-280.

Lundberg, O. (1940) On random processes and their application to sickness and accident statistics, Upsala: *Almqvist and Wiksells*.

Marshall, A.W. and Olkin, I. (1985) A family of bivariate distributions generated by the Bivariate Bernoulli distribution, *Journal of the American Statistical Association*, Vol 80, pp. 332-338.

Mellinger, G.D., Sylwester, D.L., Gaffey, W.R. and Manheimer, D.I. (1965) A mathematical model with applications to a study of accident repeatedness among children, *Journal of the American Statistical Association*, 60, pp. 1046-1059.

Mitchell, C.R. and Paulson, A.S. (1981) A new bivariate negative binomial distribution, *Naval Research Logistics Quarterly*, 28, pp. 359-374.

Mood, A.M. Graybill, F.A. and Boes, D.C. (1974) *Introduction to the Theory of Statistics*, 3rd ed., Auckland: McGraw-Hill series in probability and statistics.

Morrison, D.G. (1969) Conditional trend analysis: A model that allows for nonusers, *Journal of Marketing Research*, VI, pp. 342-346.

Morrison, D.G. and Schmittlein, D.C. (1981) Predicting future random events based on past performance, *Management Science*, Vol. 27, No. 9. pp. 1006-1022.

Mosimann, J.E. (1962) On the compound multinomial distribution, the multivariate β -distribution and correlations among proportions, *Biometrika*, 49, pp. 65-82.

Mosimann, J.E. (1975) Statistical problems of size and shape, II: Characterizations of the lognormal, gamma and Dirichlet distributions, *Statistical Distributions in Scientific Work* (Patil, G.P., Kotz, S. and Ord, J.K., editors), Dordrecht, Holland: D.Reidel Publishing Company, 1975, pp. 219-239.

Montmort, P.R.de (1713) *Essai d'analyse sur les jeux de hazards*, Chez Laurent Le Conte, Quay des Augustins à la ville de Montpellier, 2nd ed.

Newbold, E.M. (1926) A contribution to the study of the human factor in the causation of accidents, *Industrial Health Research Board*, Report No. 34, London, H.M. Stationery Office.

Patil, G.P. and Shorrock, R.W. (1966) Stochastic and sampling processes for the logarithmic series distribution. *Office of Aerospace Research, U.S. Air Force, Report ARL*, pp. 66-147.

Poisson, S.D. (1837) Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, *Précédées des Règles Générales du Calcul des Probabilités*. Paris: Bachelier, Imprimeur-Libraire pour les Mathématiques, la Physique etc.

Polya, G. (1931) Sur quelques points de la théorie des probabilités, *Ann. Inst. Henri Poincaré*, 1(2), pp. 61-117.

Rubinstein, G.Z. (1985) *Models for count data using the Sichel distribution*, PhD Thesis, University of Cape Town.

Sichel, H.S. (1951) The estimation of the parameters of a negative binomial distribution with special reference to psychological data, *Psychometrika*, Vol. 16, No. 1, pp. 107-127.

Sichel, H.S. (1971) On a family of discrete distributions particularly suited to represent long-tailed frequency data, *Proceedings of the Third Symposium on Mathematical Statistics*, (N.F. Laubscher, ed.), Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa, pp. 51-97.

Sichel, H.S. (1973a) Statistical valuation of diamondiferous deposits, *Journal of the South African Institute of Mining Metallurgy*, 73, pp. 235-243.

Sichel, H.S. (1973b) The density and size distribution of diamonds, *Bulletin of the International Statistical Institute*, 45, pp. 420-427.

Sichel, H.S. (1974) On a distribution representing sentence length in written prose, *Journal of the Royal Statistical Society, A*, 137, pp. 25-34.

Sichel, H.S. (1975) On a distribution law for word frequencies, *Journal of the American Statistical Association*, 70, pp. 542-547.

Sichel, H.S. (1982) Repeat-buying and the Generalized Inverse Gaussian-Poisson distribution, *Applied Statistics*, 31, No. 3, pp. 193-204.

Sichel, H.S. (1985) A Bibliometric distribution which really works, *Journal of the American Society for Information Science*, Vol. 36, No. 5, pp. 314-321.

Sichel, H.S. (1985) Species frequency and the Generalized Inverse Gaussian-Poisson distribution. To appear.

Skellam, J.G. (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Ser. B*, 10, pp. 257-261.

Smith, D.M. (1983) Maximum likelihood estimation of the parameters of the Beta-Binomial distribution, *Applied Statistics*, 32, pp. 196-204.

Taillie, C., Ord, J.K., Mosimann, J.E. and Patil, G.P. (1979) Chance mechanisms underlying multivariate distributions, *Statistical Distributions in Ecological Work*, (Ord, Patil and Taillie, editors), Fairland, Maryland: International Co-operative Publishing House, pp. 157-191.

Tweedie, M.C.K. (1957) Statistical properties of Inverse Gaussian distributions, I, *The Annals of Statistics*, 28, pp. 362-377.

Yule, G.U. (1944) *The Statistical Study of Literary Vocabulary*, Cambridge University Press, London.

Author Barr Aiala

Name of thesis New Statistical Models For Discrete Uni- And Multivariate Data Sets With Special Reference To The Dirichlet Multinomial Distribution. 1987

PUBLISHER:

University of the Witwatersrand, Johannesburg

©2013

LEGAL NOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed, or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.