Lilly Research Laboratories and
The Medical College of Virginia, Virginia Commonwealth Univ.

Methods of Discrimination Among Stochastic Models of the Negative Binomial Distribution with an Application to Medical Statistics[1]

F. W. ROCKHOLD and S. J. KILPATRICK

Abstract

The Negative Binomial Distribution (NBD) can be derived from several underlying stochastic processes. For the NBD to be useful in practice as a descriptive model, it would be of interest to know which underlying process it is characterizing. In this paper, methods for discriminating among three different NBD underlying processes, proneness, linear contagion, and spells, are developed and studied. The methods are based on the characteristics of the processes and employ the theory of the bivariate NBD and waiting times. The NBD has been shown to consistently fit data on the incidence of disease. Knowledge of the underlying process generating the data is important for meaningful interpretation of such findings.

Key words: *Negative binomial, stochastic processes, model discrimination, disease processes.*

Introduction

The negative binomial probability distribution (NBD) as defined by;

$$(1) \qquad P(X=x) = \frac{\Gamma(k+x)}{\Gamma(k)\,x!} \left(\frac{k}{m+k}\right)^k \left(\frac{m}{m+k}\right)^x \quad k, m > 0 \quad x = 0, 1, 2, \ldots$$

can be derived from several underlying models or processes (LOESCHKE and KOHLER, 1976 and BOSWELL and PATIL, 1970). The parameter m is such that $E(X) = m$ and k controls the skewness of the distribution. The expression for the variance is $V(X) = m(m+k)/k$. These can be easily derived from the probability generating function:

$$(2) \qquad g_x(z) = \left(\frac{k}{m+k-mz}\right)^k \quad |z| < 1 \,.$$

The parameters m and k can take on vastly different interpretations depending on the underlying model which is presumed to have generated the NBD data. Thus, for the NBD to be useful as a descriptive or predictive tool, some knowledge of

the underlying model is necessary. This paper presents methods for discriminating among several underlying models of the NBD and results of simulation studies evaluating them.

One area of application where such discrimination methods are useful is in data collected on the occurrence of illnesses in medicine. For instance, the number of episodes of illness experienced by a person in a given period of time (say a year) and recorded by a physician in the general practice of medicine has been shown to follow the negative binomial probability law (Kilpatrick, 1975; 1977; Ashford, 1972, and Spencer, 1971). An episode is defined as: "The occurrence of a specific problem or illness in a patient extending over a period of time from its onset to resolution (Froom, et al., 1977)."

The nature of data recording restricts one to studying the incidence (onset) of episodes because little or no information is available on their duration. The concept could be expanded to include, say, all physician visits, or focused on one specific disease. Although the example given relates to episodes, the methods presented herein could be applied to all such data that can be fit by the general NBD model. Examples of application are also available in the fields of psychology (Mintz, 1954) and sociology (Eaton and Fortin, 1978).

Model Development

Three of the many derivations of the NBD were selected for study as underlying models which had realistic interpretations in terms of medical statistics. The names given these three here are the *proneness*, *linear contagion*, and *spells* models.

Proneness Model

The concept of the proneness model grew, as did much of this general area, out of the field of accident statistics analysis. Greenwood and Yule (1920) and Newbold (1927) developed and studied the model in that field. In the area of medical statistics, it has been studied by Ashford and Hunt (1974) and others. The model is based on the concept that each individuals' episodes within a given period of time follow a Poisson process, with parameter λ. In turn, suppose this parameter λ is an observation from a gamma distribution. Thus, the likelihood of having an episode of illness or event is constant for the individual over time, t, but varies from individual to individual in the population. For any given λ then, the distribution of the number of events $X(t)$, in $(0, t)$ follows the Poisson law:

$$(3) \qquad P[X(t) = x \mid \Lambda = \lambda] = \frac{(\lambda t)_x e^{-\lambda t}}{x!} \quad \lambda, x, t > 0 \quad x = 0, 1, 2 \ldots$$

Now since λ is distributed as a gamma random variable with parameters α and β;

$$P_\Lambda(\lambda) = [\beta^\alpha \Gamma(\alpha)]^{-1} \lambda^{\alpha-1} e^{-\lambda/\beta}, \quad \lambda, \alpha, \beta > 0 ,$$

the probability law for the distribution of events in the entire population of individuals can be represented by:

$$(4) \qquad P[X(t) = x] = \int_0^\infty P[X(t) = x \mid \Lambda = \lambda] \, P_\Lambda(\lambda) \, d\lambda$$

$$= \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) \, x!} \left(\frac{\beta t}{1 + \beta t}\right)^x \left(\frac{1}{1 + \beta t}\right)^\alpha \quad x = 0, 1, 2. \ldots$$

Expression (4) reduces to (1) if one assumes a unit time interval and $m = \alpha\beta$ and $k = \alpha$. Thus the model allows the likelihood or "proneness" for illness to vary on a continuum from 0 to ∞. If the proneness model held, the parameters α and β could be used as an index of "health" as discussed in CHIANG (1965). One undesirable property of this model is its lack of "memory", i.e. the past occurrence of events has no effect on the likelihood of future events, this property is considered in the model outlined in the next section.

Linear Contagion Model

A second process leading to the NBD also considered by GREENWOOD and YULE (1920) is labeled here the linear contagion model. It can be thought of as a birth process with linear birthrate (PARZEN, 1962), and thus a continuous time MARKOV chain. In the present application, only the positive contagion or birth model was considered. (The negative contagion model does not lead to the NBD). In contrast to the proneness model which has no "memory", this model is conditional on the number of previous episodes occurring in the observation period. After each episode, the chance of a further episode occurring increases. The stochastic intensity function $q_j(t)$ controlling an individual's likelihood to incur an illness (the $j + 1^{\text{th}}$) in the next Δt given that j have occurred previously is given by

$$(5) \qquad q_j(t) = \nu + \gamma j \quad j = 0, 1, 2, \ldots \quad \nu, \gamma > 0 \, .$$

The assumption being made in this application of the model is that $x(0) = 0$, i.e. no events have occurred at the start of the observation period, $t = 0$. The same assumption was made under the proneness model. In general, one would want complete past information and the model can be developed for $x(0) = m > 0$, however, such data is most often unavailable. The assumption is tenable for the purpose of model discrimination.

To derive the transition probability function one must construct the KOLMOGOROV forward differential equations using (5) and then solve those using a general solution found in PARZEN (1962). Using that approach and the assumption that $X(0) = 0 \Rightarrow P[X(0) = 0] = 1$ we would arrive at;

$$(6) \qquad P[X(t) = x \mid X(0) = 0] = \frac{\Gamma(\nu/\gamma + x)}{\Gamma(\nu/\gamma) \, x!} \, e^{-\nu t} \, (1 - e^{-\gamma t})^x \quad x = 0, 1, 2, \ldots$$

which once again reduces to (1) if one assumes a unit time interval and $m = (\nu/\gamma) \times$
$\times (e^{\gamma t} - 1)$ and $k = \nu/\gamma$.

One facet of this model which causes concern is that the intensity (5) is unbounded over time. Since illness incidence data is collected over relatively short periods of time, this difficulty is not a major one for such an application of the positive linear contagion model. There is a stochastic model leading to the NBD (1) similar to linear contagion which is bounded by time; the Polya process (LUNDBERG, 1940 and FELLER, 1966). It can be seen from the POLYA intensity:

$$q_j(t) = \frac{\lambda (1 + bj)}{1 + \lambda bt}, \quad \lambda, \quad b > 0$$

that time elasping subsequent the occurrence of an event decreases the intensity, thus putting a bound on it. However, it has been shown that the methods presented herein would not be able to distinguish between this model and the proneness model (ROCKHOLD, 1978). This concept was alluded to by FELLER (1966). The deterministic properties of the two models are identical, which might thus allow one to think of the proneness model in more general terms.

Spells Model

This third model was first considered theoretically by QUENOILLE (1949) and has been applied to episode data by ASHFORD (1972) and in mental health by EATON and FORTIN (1978). The model is based on the concept that each individual experiences "spells" or groups (in time) of events or illness. The number of these spells in a given period of time is assumed to follow a POISSON process. The number of episodes occurring within a spell is then an observation from a log series distribution (FISHER, et al., 1943). The probability law of the number of episodes experienced by an individual in $(0, t)$, $X(t)$, can be derived using the theory of generalized POISSON processes (PARZEN, 1962). The probability generating function (p.g.f.) of a generalized Poisson process is

$$g_2(z) = \exp \left(\eta t \left[(g_1(z) - 1) \right] \right),$$

where η is the POISSON law parameter and $g_1(z)$ is the probability generating function of the generalizing distribution. In this case $g_1(z)$ is the p.g.f. of the log series distribution with parameter ϑ;

$$g_1(z) = \frac{\ln (1 - \vartheta z)}{\ln (1 - \vartheta)}.$$

Thus the p.g.f. of the generalized distribution is:

$$g_2(z) = \left(\frac{1 - \vartheta}{1 - \vartheta z} \right)^{-\eta t / \ln(1 - \vartheta)}$$

which is in the form of that for the NBD (2) with

$$m = -\frac{\vartheta \eta t}{(1 - \vartheta) \ln (1 - \vartheta)}$$

and

$$k = \frac{-\eta t}{\ln (1 - \vartheta)} .$$

This results in the probability function;

$$(7) \qquad P[X(t) = x] = \frac{\Gamma \left(x - \dfrac{\eta t}{\ln (1 - \vartheta)} \right)}{\Gamma \left(\dfrac{-\eta t}{\ln (1 - \vartheta)} \right) x} (1 - \vartheta)^{-\eta t / \ln(1 - \vartheta)} \vartheta^x$$

$$x = 0, 1, 2, \ldots \quad \eta > 0, \quad 0 < \vartheta < 1$$

which again reduces to (1) upon substitution for m and k and assuming a unit time interval.

In summary then, three underlying models of the NBD have been described. The proneness (Polya) model assumes a constant intensity for an episode, but this intensity varies from individual to individual in the population. The positive linear contagion model allows the intensity to change with the occurrence of an episode and this gives the process "memory". The parameters are assumed constant in the population at the start of observation. The spells model assumes that the intensity to have a spell of illness remains constant over time as well as in the population, but that the number of episodes within a spell is a random variable distributed as a log series distribution. Over an interval of time $(0, t)$ each of the above models could explain the distribution of the number of events $X(t)$ as a NBD, each with a different interpretation. Methods of discriminating among these models are now considered. The need for such methods has been pointed out by FELLER (1943). He said they were necessary to avoid making false inferences from the fit of the NBD to data, as mentioned above.

Model Discrimination Methods

The general NBD model discrimination problem has been approached by other investigators ABBOUS and KERRICH (1951), BATES and NEYMAN (1952). An attempt using some data was made by FROGGATT (1969). Although some of the methods considered herein are extensions of procedures suggested by others, the evaluation of the methods using computer simulations has not appeared at all in the literature.

The tests considered here are based on the characteristics of the underlying models themselves. All but one of the procedures are developed based on the concept that the data observation period is divided into two halves; two years for

instance. If an interval of time $(0, t)$ is split into two parts $(0, t_1)$ and $(t_1, t - t_1)$ the observed number of episodes $X(t) = x$ could be reexpressed as $X_1(t_1) = x_1$ and $X_2(t - t_1) = = X_2(t_2) = x_2$ where $x_1 + x_2 = x$ and $t_1 + t_2 = t$. Thus two variates are observed on each individual: $X_1(t_1)$ and $X_2(t_2)$ are the number of episodes experienced in the first and second half of the observation period respectively. The bivariate form of the NBD is thus considered as the model for the two intervals of time. The general form of the bivariate NBD (BNBD) assuming two equal unit intervals of time is:

$$(8) \qquad P[X_1(1) = x_1, X_2(1) = x_2]$$
$$= \frac{\Gamma(k + x_1 + x_2)}{\Gamma(k) \, x_1! x_2!} \left(\frac{k}{k + m_1 + m_2} \right)^k \left(\frac{m_1}{k + m_1 + m_2} \right)^{x_1} \left(\frac{m_2}{k + m_1 + m_2} \right)^{x_2}$$
$$x_1, x_2 = 0, 1, 2, \dots \quad m_1, m_2, k > 0$$

where k is as in (1) and m_1 and m_2 are such that $E(X)_1 = m_1$ and $E(X_2) = m_2$. The BNBD has been studied by ARBOUS and KERRICH (1951) and BATES and NEYMAN (1952). A form of (8) can be constructed for each of the underlying models considered. A review of this can be found in SIBUYA, et al. (1964).

Correlation Test

One of the characteristics of the spells model version of the BNBD is that the correlation of X_1 and X_2 is zero. This is due to the fact that the intensity to experience an illness remains constant both in time and in the population. The proneness model intensity on the other hand although constant within individuals, due to the varying intensities in the population, exhibits a positive correlation between time periods of

$$(9) \qquad \varrho = \left(\frac{m_1 m_2}{(m_1 + k)(m_2 + k)} \right)^{1/2}.$$

The fact that correlation between consecutive time periods is a characteristic of the proneness model was first proposed by NEWBOLD (1927), even though the BNBD had not yet been explicitly developed. Likewise, the linear contagion model exhibits a positive correlation which provides evidence against the spells model. The estimator chosen for measuring correlation was PEARSON's correlation coefficient, r;

$$r = \frac{\sum\limits_{i=1}^{N} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left(\sum\limits_{i=1}^{N} (x_{1i} - \bar{x}_1)^2 \sum\limits_{i=1}^{N} (x_{2i} - \bar{x}_2)^2 \right)^{1/2}}.$$

This was thought a priori to be a reasonable estimate of the population correlation because of the fact that the conditional expectation of x_2 given x_1 is a linear function of x_1, similar to the relationship in the bivariate normal. To test the hypothe-

sis $H_0 : \varrho \leqq 0$ (spells model) versus $H_a : \varrho > 0$, the following test statistic was studied:

$$(10) \qquad \sqrt{N-2} \cdot r / \sqrt{1-r^2} \sim N(0, 1) .$$

Mean Difference Test

If the hypothesis of zero correlation (spells model) is rejected using (10), one would then be interested in discriminating between the proneness and linear contagion models. It is a simple matter to show using the linear contagion form of the BNBD derived from (5) and (8), that $m_2 = e^\gamma m_1$. Since γ is constrained to be positive, this implies that $m_2 > m_1$. Similarly since the intensities remain constant over time in the proneness model, it is not surprising that $m_2 = m_1 = m$. So a statistic that would test $H_0 : m_2 \leqq m_1$ versus $H_a : m_2 > m_1$ would be of interest. Employing the central limit theorem, such a test statistic would take the form;

$$\frac{m_2 - m_1}{\sqrt{V(m_2 - m_1)}} \sim N(0, 1) .$$

An estimate of $V(m_2 - m_1)$ can be constructed as follows;

$$V(m_2 - m_1) = V(m_2) + V(m_1) - 2 \operatorname{Cov}(m_2, m_1) .$$

From the theory of the BNBD (ARBOUS and KERRICH, 1951);

$$V(m_i) = \frac{m_i(m_i + k)}{Nk} , \quad i = 1, 2,$$

and

$$\operatorname{Cov}(m_1, m_2) = \frac{m_1 m_2}{Nk} .$$

Therefore,

$$(11) \qquad V(m_2 - m_1) = \frac{1}{Nk} [m_2(m_2 + k) + m_1(m_1 + k) - 2m_1 m_2] .$$

Under the null hypotheses of the proneness model $m_1 = m_2 = m$, so (11) reduces to

$$V(m_2 - m_1) = \frac{2m}{N} .$$

The UMVUE's of m_1, m_2 and m are $\bar{x}_1$, $\bar{x}_2$, and $(\bar{x}_1 + \bar{x}_2)/2$ respectively. Thus the test statistic becomes

$$(12) \qquad \frac{\sqrt{N} \, (\bar{x}_2 - \bar{x}_1)}{\sqrt{\bar{x}_2 + \bar{x}_1}} ,$$

which is distributed asymptotically as a $N(0, 1)$ random variable.

Waiting Time Test

One additional method of indirectly testing $H_0: m_2 = m_1$ (proneness) versus $H_a: m_2 > m_1$ (linear contagion) is to consider episode occurrence times. The measure of time is the amount of time elapsing from the beginning of observation (start of data collection) to the occurrence of the event. So in that way these times could be thought of as "waiting times" in the stochastic sense. Suppose in a population of patients, the time until the jth event in the ith individual is denoted by w_{ij}, where $i = 1, \ldots, N$ and $j = 1, \ldots x_i$. Within patient i in a population of N patients, the proneness model assumes that the number of episodes, x_i, follows a POISSON law (3) with parameter λ_i. Now the waiting times in a POISSON process are distributed as order statistics from a uniform distribution (PARZEN, 1962). If the w_{ij} are converted to (0, 1) random variables by treating the observation period as a unit interval, then under the Poisson law;

$$E(w_{ij}) = \frac{1}{2} \quad \text{and} \quad V(w_{ij}) = \frac{1}{12}.$$

Under the proneness model, episodes within individuals and individuals are all independent of one another so the expected value and variance of the mean waiting time, $\bar{w}$, can be expressed as

$$E(\bar{w}) = \frac{1}{2} \quad \text{and} \quad V(\bar{w}) = \frac{1}{12n}$$

where $n = N\bar{x}$, or the total number of events occurring in the time period. With these expressions, as n (or N) increases;

$$(13) \qquad \sqrt{12n}\left(\bar{w} - \frac{1}{2}\right) \sim N(0, 1).$$

This relationship can be used to test $H_0: \bar{w} = 1/2$ versus $H_a: \bar{w} \neq 1/2$ and is similar to a procedure used by BATES (1955) to discriminate between the POISSON and another form of the linear contagion model. In the present application we are interested in testing $H_0: \bar{w} \leq 1/2$ (proneness) versus $H_a: \bar{w} > 1/2$ (positive linear contagion). The alternative hypothesis is related to the finding in the bivariate case that $m_2 > m_1$. If the two time periods used in the bivariate model are considered as one long period, then under the (positive) linear contagion model more events will occur toward the end of the interval, thus increasing the mean waiting time, $\bar{w}$. So the statistic in (13) would be used to test for the proneness versus positive linear contagion models. Simulation studies described in the next section were done to examine the behavior of the test statistics in (10), (12), and (13).

Simulation Results

The simulation study consisted of computer generating the experiences of a patient population of 1000 for each of the three NBD underlying models studied. Each simulated patients' period of observation was analogous to two years. There

were 200 repetitions at $N = 1000$ of this procedure, each at 5 combinations of k and m_1. These combinations were chosen based on results from the analysis of data from the Virginia Family Practice Data System (ROCKHOLD and KILPATRICK, 1979) and are listed in Table 1 along with the translation in terms of the parameters of each of the underlying models.

The simulation results concerning the test of $H_0 : \varrho \leqq 0$ versus $H_a : \varrho > 0$ are considered first. The results in Table 2 indicate that the sample variance of

Table 1

NBD model discrimination simulation study. Bivariate NBD parameter combinations used in the simulations and their translation into the parameters of the three models under study

Case	General BNBD		Proneness		Spells		Linear Contagion	
	k	m_1	α	β	η	θ	ν	γ
1	2.0	1.0	2,0	0.75	1.12	.43	1.12	.56
2	1.5	2.0	1.5	1.33	1.27	.57	1.27	.85
3	2.0	2.0	2.0	1.00	1.39	.50	1.39	.69
4	2.0	2.5	2.0	1.25	1.62	.56	1.62	.81
5	1.5	1.0	1.5	0.67	0.77	.40	0.77	.51

Table 2

Results of the simulation study on the correlation test for NBD model discrimination
Theoretical type I error probability $= .05$. Theoretical variance of $r = .001$
Results of 200 simulated distributions at $N = 1000$

Case	Sample Variance of r	Type I Error Rate for Test of $H_0 : \varrho \leqq 0$ vs $H_a : \varrho > 0$
1	.00093	.050
2	.00109	.040
3	.00113	.060
4	.00103	.045
5	.00094	.040

PEARSON's r was very close to the theoretical value of .001 under the null hypothesis of the spells model. Table 2 also lists the estimated Type I error probabilities for the statistic in (10) for the five parameter combinations at a Type I error level of .05. All of the estimated Type I error probabilities are at or near the theoretical value.

The second method considered involved testing $H_0 : m_2 \leqq m_1$ (proneness) versus $H_a : m_2 > m_1$ (linear contagion) using (12). Table 3 lists results which indicate that the variance estimator of $(\bar{x}_1 + \bar{x}_2)/N$, for V $(m_2 - m_1)$ under the null hypothesis of proneness, is a good one based on the comparison to the sample variance of the simulated differences. The hypothesis test results in Table 3 indicate that at all of the parameter combinations the estimated Type I error probabilities are near the expected value of .05, with the exception of $k = 1.5$ and $m_1 = 1.0$ where the value of .09 was higher than desired.

Table 3

Results of the simulation study on the mean difference test for NBD model discrimination
Theoretical type 1 error probability = .05
Results of 200 simulations at $N = 1000$

Case	Theoretical Variance of $m_2 - m_1$	Sample Variance of $m_2 - m_1$	Type I Error Rates for Test of H_0: $m_2 \leqq m_1$ vs H_a: $m_2 > m_1$
1	.00294	.00327	.050
2	.00388	.00362	.055
3	.00388	.00363	.035
4	.00483	.00432	.030
5	.00197	.00241	.090

Table 4

Results of the simulation study on the waiting time test for NBD model discrimination
Theoretical type I error probability = .05
Theoretical mean waiting time, $\bar{w} = .500$
Theoretical variance, $\bar{w} = .08333$
Results of 200 simulations at $N = 1000$

Case	Sample Mean of $\bar{w}$	Sample Variance of $\bar{w}$	Type I Error Rates for Test of H_0: $\bar{w} \leqq 1/2$ vs H_a: $\bar{w} > 1/2$
1	.500	.08320	.070
2	.500	.08328	.035
3	.500	.08335	.045
4	.499	.08339	.020
5	.500	.08339	.070

The test for H_0: $\bar{w} \leq 1/2$ (proneness) versus H_a: $\bar{w} > 1/2$ (linear contagion) using (13) involves the mean and variance of the waiting times w_{ij}. Table 4 reveals that under the proneness model, the assumptions of $E(w_{ij}) = 1/2$ and $V(w_{ij}) = 1/12$ are reasonable. The results in Table 4 show that in addition the statistic in (13) has estimated Type I error probabilities close to the expected value of .05.

Power of the Tests

The evaluation of the power of these tests is very involved, and is the subject of a separate paper. However, the empirical evidence obtained from simulation studies imply that the power at the sample size used and the alternatives indicated by parameter values chosen is more than adequate.

Summary and Conclusions

Given that data, in particular recorded incidence of episodes of illness, can be modelled by the NBD (1), methods have been presented for determining the underlying model generating the data. Simulation studies showed these methods to be

reliable. Certainly other models of·the NBD could be considered which might lead to the development of additional tests. In addition, work on other parameter combinations and smaller sample sizes should be done.

An attempt has been made to provide methods of interpreting the parameters of the NBD when it is used as an empirical model. Hopefully this will lead to practical application of such a model as opposed to treating it as a mathematical artifact.

Acknowledgements

Résumé

La distribution binomiale negative (DBN) peut etre derivee de plusieurs procedes stochastiques communs. Pour que la (DBN) soit utile et practique comme modele descriptif, il serait interessant de savoir quel prodece commun elle caracterise.

Dans cette etude des methodes de descrimination entre trois differents procedes communs tels que la predisposition, la contagion lineaire et les crises, sont developpees et etudiees. Les methodes sont basees sur les caracteristiques des procedes et emploient la theorie de la (DBN) bivariable et les periodes d'attente. La (DBN) a demontre qu'elle s'accordait tous jours avec l'incidence de la maladie. Pour obtenir une interpretation significante de tels resultats, il est important de connaitre le procede commun fournissant les donnees.

References

ARBOUS, A. G. and KERRICH, J. E., 1951: Accident Statistics and the Concept of Accident Proneness. Biometrics 7, 340–432.

ARBOUS, A. G. and SICHEL, H. S., 1954: New Techniques for the Analysis of Accident Data. Biometrika 41, 77–90.

ASHFORD, J. R., 1972: Patient Contacts in the National Health Service. The Statistician 21 (4), 265–289.

ASHFORD, J. R., and HUNT, R. G., 1974: The Distribution of Doctor Patient Contacts in the National Health Service. J. Roy Statist. Soc. A, 137 (3), 347–371.

BATES, G. E., 1955: Joint Distribution of Time Intervals for the Occurence of Successive Accidents in a Generalized Poyla Schema. Ann. Math. Statist. 26, 705–720.

BATES, G. E. and NEYMAN, J., 1952: Contributions to the Theory of Accident Proneness. I. An Optimistic Model of the Correlation between Light and Severe Accidents. II. True or False Contagion. U. Cal. Publ. in Statistics (1) 10, 215–275

BISHOP, Y. M. M., FEINBERG, S. E., and HOLLAND, P. W., 1975: Discrete Multivariate Analysis. The MIT Press.

BOSWELL, M. T. and PATIL, G. P.: Chance Mechanisms Generating the Negative Binomial Distributions from *Random Counts in Scientific Work*, Vol. 1, 3–22, G. P. Patil, Ed. Penn. State University Press.

CHIANG, C. L., 1965: An Index of Health: Mathematical Models. Vital and Health Statistics, Series 2, No. 5, DHEW, USA.

EATON, W. W. and FORTIN, A., 1978: A Third Interpretation for the Generating Process of the Negative Binomial Distribution. Amer. Soc. Rev. 43, 264–267.

FELLER, W., 1943: On a General Class of "Contagious Distributions". Ann. Math. Statist. 14, 389–400.

FELLER, W., 1966: An indroduction to Probability Theory and Its Applications, Vol. II, John Wiley & Sons, New York.

FISHER, R. A., CORBETT, S. A., and WILLIAMS, C. V., 1943: The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. J. Animal Ecology **12**, 42–58.

FROOM, J., ANDERSON, J. E., BASS, M., BOYLE, R. M., FREEMAN, D. M., KILPATRICK, S. J., and WOOD, M. A Glossary for Primary Care. J. Fam. Pract. (5) **4**: 633–638.

GREENWOOD, M. and YULE, G. U., 1920: An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Paticular Reference to the Occurrence of Multiple Attacks of Disease of of Repeated Accidents. J. Roy. Statist. Soc. A. **83**, 255–279.

KILPATRICK, S. J., 1975: The Distribution of Episodes of Illness: A Research Tool in General Practice. J. Roy. Coll. Gen. Pract. **25**, 686–690.

KILPATRICK, S. J., 1977: An Empirical Study of the Distribution of Illness Recorded in the 1970–71 National Morbidity Survey. Appl. Statistics **26** (1), 26–33.

LOESCHKE, V. and KOHLER, W., 1976: Deterministic and Stochastic Models of the Negative Binomial Distribution and the Analysis of Chromosomal Aberrations in Human Leukocytes. Biom. Z. **18**, 427–451.

LUNDBERG, O., 1940: On Random Processes and Their Application to Stickness and Accident Statistics. Alhmquist and Wijks.

MINTZ, A., 1954: The Inference of Accident Liability from the Accident Record. J. Appl. Psych. **38** (6), 41–46.

NEWBOLD, E. M., 1927: Practical Applications of the Statistics of Repeated Events, Particularly to Industrial Accidents. J. Roy. Statist. Soc. **90**, 487–547.

PARZEN, E., 1962: Stochastic Processes. Holden-Day.

QUENOUILLE, M. H., 1949: A Relation Between the Log Series, Poisson Distribution, and Negative Binomial. Biometrics **5**, 162–164.

ROCKHOLD, F. W., 1978: Discrimination Among Stochastic Models of the Nagative Binomial: An Application to Episodes of Illness. Ph. D. Dissertation, Department of Biostatistics, Medical College of Virginia/Virginia Commonwealth University, Richmond, Virginia, USA.

ROCKHOLD, F. W. and KILPATRICK, S. J. Negative Binomial Models of Demand for Primary Care in the United States. Published in the *Conference Proceedings of the North American Primary Care Research Group* (NAPCRG) Annual Meeting, Seattle, Washington, April 1979.

SIBUYA, M., IASO, Y., and SHIMZU, R., 1964: Negative Multinomial Distribution. Ann. Inst. Statist. Math. Tokyo **16**, 409–426.

SPENCER, P., 1971: General Practice and Models of the Referral Process. Health Report No. 6. Institute of Operations Research Tavistock, London.

Author's address:

FRANK W. Rockhold, Ph. D.
Senior Statistician
Statistical and Mathematical Services
Lilly Research Laboratories
307 E. McCarty Street
Indianapolis, Indiana 46285 U.S.A.