VOL. XLIX, Part I No. 91

PROCEEDINGS

MAY 21, 22 and 23, 1962

AN INTRODUCTION TO THE NEGATIVE BINOMIAL DISTRIBUTION AND ITS APPLICATIONS

BY

LEROY J. SIMON

I. OBJECTIVE

The description, interpretation, and curve fitting of the negative binomial distribution has become a topic of great interest to American acturaries in the last few years. What is it? Where did it come from? What does it mean? How can it be used? These and many other questions have been asked by all of us. The first thing to do, of course, is to check the textbooks and references in our personal libraries. After this has been done, some questions may still remain or some new ones brought to mind.

The purpose of this paper is to present a bibliography selected especially for actuaries from the hundreds of papers and texts that deal with the subject, to organize the material, and to make a few comments on each reference so that the interested reader may choose those references which are of particular interest to him and study them in detail. Two mechanical models are also described which may be used by the reader to actually generate negative binomial distributions.

II. THE FUNDAMENTALS

If only one paper can be read, it should be Arbous and Kerrich.[1] The first part is a critical evaluation and gives a good description in non-mathematical terms of the accident proneness concept. The significant literature on the subject up to 1951 is reviewed. These authors force us to separate the concepts of accident proneness and accident liability. Accident proneness is an attribute of the individual which does not fluctuate and does not tend to increase or decrease over a long period of time. Accident liability includes accident proneness plus the effects of age, experience, fatigue, emotional state, general health, and so forth. In the mathematical section of the paper the negative binomial is developed and discussed from two viewpoints. If

[1] See Appendix A "Selected Bibliography" for complete citation of this and other references throughout the paper.

we assume that accidents happen in the form of a Poisson distribution for each isohazardous group, and that these groups appear in the population with a relative frequency which is distributed in the form of a Pearson Type III curve, then the distribution of the number of persons having 0, 1, 2, ... accidents will be in the shape of a mathematical curve which has been given the name "negative binomial distribution." A second method for deriving this distribution is to assume that everyone starts out with the same propensity toward having an accident which remains constant until an accident occurs. When an accident does occur, the future probability for that individual is changed. This development also leads to a negative binomial curve. The paper ends with a very good discussion of the bivariate negative binomial and its relationship to accident proneness. More will be said about this approach in Section VI of this paper.

A somewhat shorter presentation and review of the field has been made by Fitzpatrick. This paper gives a good, compact, and essentially non-mathematical description of various work that has been done on the accident susceptibility problem. It also contains an extensive list of references to papers which have utilized this curve.

Kendall and Stuart develop the negative binomial in two ways. The more interesting method is in discussing sequential sampling when the objective is to continue sampling until a certain number of successes has been achieved. The number of items sampled will then follow a negative binomial distribution.

In insurance terminology, Bichsel develops the negative binomial using the Poisson and Pearson Type III curves. Although the paper is in French,[2] the mathematical development can be readily followed. The insurance data are of special interest, being based on automobile accidents occuring to cars insured by a Swiss company. The conclusions drawn by Bichsel do not coincide completely with current American practices, but seem to follow from the limited sample and his very conservative assumptions on safety factors.

The development most familiar to readers of this journal is that of Dropkin (1959). This paper again uses the Poisson and Pearson Type III approach and gives examples of curve fitting to California automobile accident data. The paper also discusses the overlap between various subgroups in the study.

In an attempt to bridge the gap of intuitive feel for the negative binomial, Simon (1960) discussed the curve using the more familiar Poisson as a referent.

III. EARLY ORIGINS

Greenwood and Yule presented the basic paper which developed the theory into a mathematical model and tested it on actual accident data. This paper is a classic and is referred to by many authors. For example, Kendall and Stuart (referred to in Section II) summarize this paper very well and give the data that Greenwood and Yule used.

During the ensuing years, a number of authors followed this 1920 development and utilized this curve in describing accident phenomena. They dealt primarily with industrial accidents and were concerned, in many cases, with the psychological and sociological problems connected with the accident proneness phenomenon. If different people had a different accident prone-

[2] I have a few copies of an English translation and permission has been obtained for limited distribution.

ness, we might improve safety if we could detect it or if we could change it.

The first work in actuarial literature that has come to my attention involving the negative binomial was by Keffer in 1929 in connection with a group life experience rating plan. He developed the theory in relationship to the relative dispersion of loss ratios about their true mean. In replying to the written discussion which followed the paper he developed the equations for the mean and variance and commented on the fact that the variance exceeded that of the Poisson distribution in a manner which he interpreted to indicate the heterogeneity of the data.

A. L. Bailey first utilized the negative binomial in the Proceedings of the Casualty Actuarial Society in 1950. He compounded the Poisson with a Pearson Type III as one of the special cases in his presentation. Although the curve is referred to variously as a negative binomial, compound Poisson, contagious, Polya-Eggenberger or an accident proneness distribution, neither Mr. Bailey nor Mr. Keffer used any of these terms in their papers.

IV. APPLICATIONS

There are numerous applications of the negative binomial in the literature. Almost all of the previously mentioned papers contain one or more examples as part of the paper itself.

Bliss presents twenty-two frequency distributions of biological data and fits negative binomials to them. The paper is excellent for many reasons in addition to the data presented. First, it gives a clear explanation of the curve using the (positive) binomial distribution as a starting point. Then three methods of fitting the curve are presented: (a) using the method of moments and the mean and variance of the observed data, (b) using a very straightforward method based on the mean and the number of zero cases, (c) using the method of maximum likelihood. Two methods are discussed for testing the goodness of fit (1) the usual χ^2 and (2) a test of the third moment of the sample compared to the value predicted from the first two moments. This is of particular interest when the tail of the curve is rather short as we often find it in insurance data. Finally, a rather unusual method of developing the negative binomial is illustrated wherein the number of bacterial colonies per microscopic field follows the Poisson distribution in repeated sampling while the number of bacteria per colony follows a logarithmic distribution. In combination, the number of bacteria per field follows the negative binomial.

Another non-insurance application is by Wise who considers a quality control problem. It was possible to assume that defects occurred at random and in a Poisson manner in each batch of the material to be sampled. Different batches had a different expectation of the mean number of defects. These two facts were compounded to produce a negative binomial distribution which was used to establish the quality control limits for the process.

A very thorough study of motor accidents by Häkkinen was done as a doctoral dissertation. Not only does he comment on the mathematical aspects, but he also goes into a number of intelligence, mechanical aptitude, coordination and psychomotor tests in an attempt to isolate specific factors which lead to higher accident rates among certain individuals.

Finally, the papers of Delaporte and Thyrion, although written in French, are still easy to follow in the mathematical developments and present inter-

esting data. In his review of the latter paper, Beard almost casually produces three bivariate negative binomial distributions (see Section VI for further discussion).

V. MODELS

It is often helpful to have some type of operating model to assist in understanding a mathematical formula. To further assist in getting an intuitive feeling for the negative binomial, two simple models have been devised. The first is to throw a six-sided die numbered as usual, and count the number of throws needed to produce six successes where a success will be defined as a "1", "2", "3" or "4" appearing face up. The distribution resulting from repeated trials of this experiment will be in the form of a negative binomial distribution. This model is suggested by the mathematical development of Feller (1957). In an experiment involving this model, about 7300 throws of a die were made and a total of six successes was achieved 809 times. The distribution is shown in Appendix B along with the theoretical expectation.

A more elaborate model was constructed to create something that may be easier to visualize as an insurance situation. Rule off a sheet of paper into S squares. Get D paper disks such as the disks punched out by an ordinary paper punch. The diameter of the disks should be small compared to the length of one side of the square. Drop the disks one at a time from a sufficient height to negate any tendency to clustering and record the distribution of the number of squares having 0, 1, 2, . . . disks on them. The resulting distribution will be in the form of the Poisson distribution. Conduct two separate experiments of this nature, the first involving 361 squares and 36 disks (representing a large group of insureds with an accident frequency of .100), and a second experiment with 49 squares and 31 disks (representing a small group of insureds with the high loss frequency of .633). Combine the results of each subsample into s single combined sample. The actual results of such an experiment conducted ten times by the author is shown in Appendix C. In repeated sampling the distribution will tend toward the probabilities shown in the theoretical column. Finally, the last column indicates the unusually close agreement with the well-known California data, which appears in Dropkin (1959) page 174. This model was suggested by David.

VI. ADVANCED TOPICS

The property and casualty actuary may find a number of extended uses of the negative binomial distribution. To properly capitalize on these, however, will require more advanced study. Some extensions of Dropkin's original paper have been made already in our Proceedings. Dropkin (1960) introduces the time element specifically into the formulas and discusses the distribution of accidents in subsequent years, given the number of accidents in some previous time period. Hewitt (1960, pp. 55-65) additionally develops expectations for the claim frequency next year, given that the person has been claim-free for 0, 1, 2 or 3 or more years. He then gets close fits to actual Canadian automobile statistics with these formulas. Simon (1961) discussed the problem of truncated distributions. In insurance this might arise when the number of claim-free insureds is not available but the distribution of poli-

cies having claims can be obtained. This paper also discusses the maximum likelihood method in fitting negative binomial curves.

An interesting and different extension was made by Hewitt (1960, pp. 41-54) when he considered the problem of mortality curve fitting over the entire life range. The negative binomial was utilized here as one of the factors.

In a highly developed mathematical-statistical presentation, Lundberg sets forth the basic tenets of random processes. The first chapter is of particular interest in its lucid description of the Polya-Eggenberger distribution through the use of an urn model where each time a black ball is drawn, a number of black balls is added to the urn and each time a white ball is drawn a number of white balls is added to the urn. This approach is then extended to a concept of a continuous set of samplings by the passage to the limit in such a fashion that (1) the number of drawings times the initial probability of success remains constant in the manner of the Poisson limit and (2) the number of drawings times the proportion of colored balls added after each drawing remains constant (i.e., there is a continuous flow of change in the probability of success as time progresses). Finally, the accident proneness approach using the Pearson Type III is shown and the two developments are demonstrated to be identical in their resulting distributions. The last chapter of the book applies the theories to accident and sickness data on the number of claims made by the same individual.

In a strictly mathematical development, Feller (1943) ties together developments by a number of authors and presents a general distribution function for combining a Poisson with any other desired distribution for inherent hazard. He then shows conditions which lead to the Polya-Eggenberger distribution and the contagious distribution. The nice, general way in which this subject is presented makes the paper valuable reading.

A very interesting approach to the analysis of data is given by Mintz when he studies the elapsed time between successive accidents. His purpose is to see if there is an indication that the time interval between successive accidents decreases with each accident. If it did, he would conclude that accident susceptibility increased for the individual with each successive accident. Conversely, if the time interval tended to increase, he would conclude that having an accident decreases the future accident susceptibility. He did not observe either effect and therefore concluded that we should retain the theory of proneness and that the individual's proneness remains relatively constant. The study was based on one year's experience of taxi drivers. There is reason to suspect that a longer period and a study of the car (under a private passenger insure-the-car automobile policy), rather than the individual driver, would probably show that inherent hazard of the car did not remain nearly so constant as the proneness of the individual. However, that is a different problem from the one being considered by Mintz, and does not detract from that work.

Bartlett's development of the negative binomial through reference to Markov chains is thought-provoking. He develops it as a "birth" process, and assuming the transition probability from one state to the next in the Markov chain increases linearly in proportion to the state that the process has reached. In accident terminology, it means that the chance of an additional accident (the next step in the Markov chain) depends upon the number which have already occurred (that is, the state the process has reached). This is the

"contagion" effect and acts as a good reminder to actuaries that the negative binomial may be developed by a multiplicity of methods, only one of which is the accident proneness approach.

The bivariate negative binomial provides one of the more interesting topics for advanced study by actuaries. As previously mentioned, Arbous and Kerrich close their paper with a discussion of this two-dimensional curve. It simply means, of course, that if you take two different periods of time and tabulate, in a two-way table, the accident experience of a group that has a bivariate negative binomial distribution, each row and each column will be distributed as a (univariate) negative binomial. This approach is particularly appropriate to insurance where we are classifying people on the basis of past experience and then predicting (through rate differentials) what the future experience in these various groups will be.

A particularly startling realization of what might happen in a classification system similar to our insurance approach was given by Maritz. He fits a Poisson distribution to univariate data, but then shows that its bivariate distribution has a marked correlation between period one and period two. This serves as a warning that even though the marginal distributions may be Poisson, there may still be a significant and marked correlation in the data. Maritz then demonstrates how a bivariate distribution may have marginal distributions of the negative binomial form, but still be absolutely useless in predicting the results in period two based on the results from period one. We have all recognized this in insurance when we have emphasized that the rate differentials in something like the Safe Driver Insurance Plan *must* be based on the actual experience developed by the various classes otherwise they may possibly be completely fictitious and unfairly discriminatory differentials.

Edwards and Gurland present a rather intricate concept in a clear and careful manner. First they discuss the bivariate negative binomial. They then comment on a correlated bivariate Poisson, in which there is a correlation between the number of accidents in one period and the number in the other; but the marginal distributions are each Poissons. As a final step, they compound the correlated bivariate Poisson with a Pearson Type III curve. The resultant distribution thus incorporates the concepts of both the negative binomial and a correlation between different time periods.

The books and papers included in this review are necessarily only a few of the many references to the negative binomial in the literature. They were chosen to show the variety of uses of the distribution and to appeal particularly to actuaries. There is a great deal of exploration and application yet to do and I hope our Proceedings will contain much of the good work in the future.

APPENDIX A

SELECTED BIBLIOGRAPHY

I. Objective

II. The Fundamentals

Arbous, A. G. and Kerrich, J. E., "Accident statistics and the concept of accident-proneness. Part I: A critical evaluation. Part II: The mathematical background," *Biometrics* 7:340-429, 1951

Bichsel, F., "Une méthode pour calculer une ristourne adéquate pour années sans sinistres." (A method of calculating an adequate no-claim bonus for years without losses.) *The Astin Bulletin* 1:106-122, 1960

Dropkin, L. B., "Some considerations on automobile rating systems utilizing individual driving records," *Proceedings of the Casualty Actuarial Society* 46:165-176, 1959, and discussion by R. A. Bailey, *PCAS* 47:52-56, 1960

Fitzpatrick, R., "The detection of individual differences in accident susceptibility," *Biometrics* 14:50-66, 1958

Kendall, M. G. and Stuart, A., *The advanced theory of statistics, Vol. 1.* New York: Hafner Publishing Co., pp. 129, 225, 1958

Simon, L. J., "Negative binomial and Poisson distributions compared," *Proceedings of the Casualty Actuarial Society* 47:20-4, 1960

III. Early Origins

Bailey, A. L., "Credibility procedures—La Place's generalization of Bayes' rule and the combination of collateral knowledge with observed data," *Proceedings of the Casualty Actuarial Society* 37:7-23, 1950

Greenwood, M. and Yule, G. Udny, "An inquiry 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* 83:255-279, 1920

Keffer, R., "An experience rating formula," *Transactions of the Actuarial Society of America* 30:130-39, 1929. See also discussion, 593-611

IV. Applications

Bliss, C. I., "Fitting the negative binomial distribution to biological data," *Biometrics* 9:176-200, 1953

Delaporte, P., "Un problème de tarification de l'assurance accidents d'automobiles examiné par la statistique mathématique," *International Congress of Actuaries* 2:121-135, 1960

Häkkinen, Sauli, "Traffic accidents and driver characteristics," Finland's Institute of Technology, Scientific Researches No. 13, Helsinki, 1958

Thyrion, P., "Etude de la loi de probabilité de la variable 'nombre de sinistres' dans l'assurance automobile," *International Congress of Actuaries* 2:25-36, 1960, and discussion by R. E. Beard, *ICA* 3:213-4, 1960

Wise, M. E., "The use of the negative binomial distribution in an industrial sampling problem," *Supplement* to *Journal of the Royal Statistical Society* 8:202-11, 1946

V. Models

David, F. N., *Probability theory for statistical methods.* London: Cambridge University Press, p. 66, 1949

Feller, W., *An introduction to probability theory and its applications, Vol. 1.* New York: John Wiley & Sons, Inc., p. 155, 1957

VI. Advanced Topics

Bartlett, M. S., *An introduction to stochastic processes.* London: Cambridge University Press, pp. 55-6, 1955

Dropkin, L. B., "Automobile merit rating and inverse probabilities," *Proceedings of the Casualty Actuarial Society* 47:37-40, 1960

Edwards, C. B. and Gurland, J., "A class of distributions applicable to accidents," *Journal of the American Statistical Association* 56:503-17, 1961

Feller, W., "On a general class of contagious distributions," *Annals of Mathematical Statistics* 14:389-400, 1943

Hewitt, C. C., Jr., "The negative binomial applied to the Canadian merit rating plan for individual automobile risks," *Proceedings of the Casualty Actuarial Society* 47:55-65, 1960

Hewitt, C. C., Jr., "A new approach to infant and juvenile mortality," *Proceedings of the Casualty Actuarial Society* 47:41-54, 1960

Lundberg, O., *On random processes and their application to sickness and accident statistics.* Uppsala, 1940

Maritz, J. S., "On the validity of inferences drawn from the fitting of Poisson and negative binomial distributions to observed accident data," *Psychological Bulletin* 47:434-443, 1950

Mintz, A., "A methodological note on time intervals between consecutive accidents," *Journal of Applied Psychology* 40:189-191, 1956

Simon, L. J., "Fitting negative binomial distributions by the method of maximum likelihood," *Proceedings of the Casualty Actuarial Society* 48:45-53, 1961

APPENDIX B

Distribution of the Number of Throws of a Die Necessary to Get Six Successes Where the Probability of Success is Two-Thirds

Number of Throws	*Actual Results*	*Theoretical Results*
6	54	71.0
7	149	142.1
8	163	165.7
9	160	147.3
10	110	110.5
11	72	73.6
12	39	45.0
13	31	25.8
14	14	13.9
15	4	7.2
16	8	3.6
17	1	1.8
18	2	.8
19	1	.4
20	1	.2
21	0	.0
22	0	.1
	809	809.0

APPENDIX C

Combined Results of Two Independent Poisson Distributions in which $m_1 = 36/361$, $N_1 = 361$, $m_2 = 31/49$ and $N_2 = 49$

Number of Accidents	*Actual Sample Results*		*Theoretical Probability*	*California Data*
	Number	*Probability*		
0	3530	.8610	.8604	.8607
1	492	.1200	.1196	.1191
2	60	.0146	.0167	.0171
3	15	.0037	.0028	.0026
4	2	.0005	.0004	.0004
5	1	.0002	.0001	.0001