Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2016; 2(3): 130-5.

Original Article

A generalization of the negative binomial distribution

Maryam Nazemipour[1], Mahmood Mahmoudi[2*]

[1] Department of Epidemiology and Biostatistics, School of Public Health, International Campus, Tehran University of Medical Sciences, Tehran, Iran
[2] Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Received 09.08.2016
Revised 21.02.2016
Accepted 17.05.2016
Published 27.08.2016

Key words:
Negative binomial;
Generalized Poisson;
Gamma distribution;
Compound distribution

ABSTRACT

Background & Aim: Consider a sequence of independent Bernoulli trials with *p* denoting the probability of success at each trial. With this definition, the probability that the n^{th} success proceed by *r* failures follows the negative binomial distribution (NB). NB model has been derived from two different forms. At first, the NB can be thought as a Poisson-gamma mixture. The second form of the NB can be derived as a full member of a single parameter exponential family distribution, and therefore considered as a GLM (generalized linear models).

Methods & Materials: We have described a new generalized NB (GNB) distribution with three parameters α, β and k obtained as a compound form of the generalized Poisson and gamma distributions. This distribution gives a very close fit for a large number of data and provides an appropriate model for numerous studies. The most important feature of this model is, its time dependent probabilities, and also it can be used for a variety of researches especially in the survival analysis.

Results: This model has been illustrated with two datasets that are indirect measures of illness, along comparing the results of the fitting with NB. Results indicate too much satisfaction. Expected frequencies have been calculated for these data sets to show that the distribution provides a very satisfactory fit in different situations.

Conclusion: Using GNB models allows analyzing very complex data. This distribution gives a very close fit for a large number of data and provides an appropriate model for numerous studies. With k = 0 the model becomes the ordinary NB and with $\alpha = 1$, it becomes a new model which we call it the generalized geometric distribution with two parameters. The most important feature of this model is its time-dependent probabilities.

Introduction

Consider a sequence of independent Bernoulli trials with *p* denoting the probability of success in each trial. With this definition, the probability that the n^{th} success is proceeded by exactly *r* failures and will be as follows:

$$\binom{n+r-1}{r} p^n q^r \qquad r = 0,1,2,\ldots \quad (0.0)$$

Hence, the distribution of the random variable X as the number of failure before the nth success follows the negative binomial distribution (NB).

In 1920, Greenwood and Yule assumed that if the parameters of the Poisson distribution followed gamma distribution, the result could be the NB and it could have a lot of applications in many situations (1). For more detail about the distribution and its properties, we can refer to the articles written by Johnson and Kotz (2) also Takacs (3). Afterward, Mohanty in 1966 considered a series of independent trials with the same probability of success (*p*) at each trial and

* Corresponding Author: Mahmood Mahmoudi, Postal Address: Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, P.O. Box: 14155-6446, Tehran, Iran. Email: mahmoodim@tums.ac.ir

Please cite this article in press as: Nazemipour M, Mahmoudi M. A generalization of the negative binomial distribution. *J Biostat Epidemiol. 2016; 2(3): 130-5*

showed number of failure *x* for encountering exactly n+βx successes has the following distribution:

$$\frac{n}{n+(\beta+1)^x}\binom{n+(\beta+1)^x}{X}P^{n+\beta x}(1-p)^x \quad (0.0)$$

Where $\beta \geq 0$ and $n > 0$ (4).

A slightly different from what was defined; Jain and Consul in 1971 introduced a generalized NB (GNB) distribution as follows:

$$b_\beta(x,n,\alpha) = \frac{n\Gamma(n+\beta x)}{x!\Gamma(n+\beta x-x+1)}\alpha^x(1-\alpha)^{(n+\beta x-x)} \quad (0.0)$$

Where

$0 < \alpha < 1 \qquad |\alpha\beta| < 1$

In this formula, the parameter β cannot be an integer (5).

Relying on the previous works of Cameron and Trivedi (6) in 1998, Winkelmann and Zimmermann (7) in 1995 discussed what was called a generalized event count model. Moreover, in this context, Greene (8) in 2006 introduced a more flexible model and called it a general NB with three parameter

The GNB distribution: Completing what was already mentioned about Greenwood and Yule, they used NB in problems of industrial accidents. It was used by Brass (9) in 1958 for human population distributions and in many other situation too. They assumed compound poison and gamma distributions in their works.

There are many situations where individuals are exposed to the continuous risk of occurrence of an event. It is after this that immunity takes place during the study period following the experience of the event in question, such that if someone experiences the event, he will remain immune because of experiencing that event again in that period. Therefore, individuals are immune to that event and it is like after its actual occurrence. For instance in problems of industrial accidents a worker might be reckoned to be continuously exposed to the risk of occurrence an accident. In such cases, the incidence of the accident is followed by a period of immunity. Or in the birth distribution, a live birth is followed by a period of non-susceptibility.

The probability of observing *x* accidents or births in a given period of time has been given by Dandekar in 1953 as a modified Poisson distribution as follows:

$$e^{-(1-kx)\lambda}\sum_{S=0}^{X}\frac{\{(1-kx)\lambda\}^S}{S!} \quad (0.0)$$

Which are the first (x + 1) terms in the Poisson series with $\mu = (1-kx)\lambda$ (10).

The modified Poisson distribution in industrial accidents issues gives the probability of meeting *x* accidents in a given period with an infinitesimal chance of an accident at every instant, with this condition that if an individual actually meet the accident at any instant, he or she remains immune to any further accidents for a subsequent fraction k of the total period under study.

Later, a slightly different from what was said, Consul and Jain defined generalized Poisson distribution as follows (11):

$$p(x) = \frac{\lambda_1(\lambda_1+x\lambda_2)^{x-1}e^{-(\lambda_1+x\lambda_2)}}{x!} \quad (0.0)$$

If we assume $\lambda_1 = \lambda$ and $\lambda_2 = -k\lambda$ also we assume the parameter of λ varies with the gamma distribution as follows:

$$\text{□}(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \quad (0.0)$$

Then we will have,

$$b_x(\alpha,\beta,k) = \int_0^\infty f(x,\lambda)\text{□}(\lambda)d\lambda = \int_0^\infty \frac{e^{-(1-kx)^\lambda}\{(1-kx)\}^x}{(1-kx)x!}\frac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)}d\lambda = \frac{1}{1-kx}\frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{1-kx}{1-kx+\beta}\right)^x\left(\frac{\beta}{1-kx+\beta}\right)^\alpha \quad (0.0)$$

Which can be considered as a generalized GNB.

This distribution somehow can be taken as a representative of many families of probability distributions, depending on the constants values of α, β and k so that if k = 0 then it will become the NB distribution, and if α = 1 then it will be a new distribution which can be called a generalized geometric distribution and it can be written as follows:

$$G(x,\beta,k) = \frac{1}{1-kx}\left(\frac{\beta}{1-kx+\beta}\right)\left(\frac{1-kx}{1-kx+\beta}\right)^x \quad (0.0)$$

Graphical representation of the distribution: To study the behavior of the GNB distribution

by varying the values of α, β and k, the probabilities for different values of x have been calculated and some graphs have been plotted for various values of this three parameters. Some of these graphs are shown in figure 1.

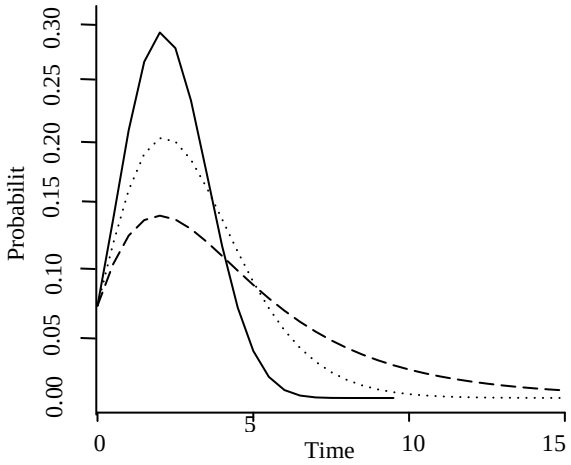

Figure 1: The graphs for fixed values of α = 9, β = 3 and with different values of k = −0.1, 0, 0.1 that has been depicted by dashed, dotted and solid lines, respectively

As we can see, when k is positive and α and β are constant, mean and variance of the distribution are smaller than the mean and variance of the NB and the bell-shaped form of it becomes narrower in compare with the NB. In addition, while k is negative and α and β are constant, mean and variance of the distribution are bigger than the mean and variance of the NB and the bell-shaped form becomes flatter.

In the second figure, when k is constant and equals to 0.1, the graph is L-shaped as long as α and β are similar, whereas it changes gradually to a bell-shaped when α becomes bigger than β.

Mean and variance of the distribution: Determination of mean and variance of the distribution is straightforward.

If m_r be the r^{th} moment of the modified Poisson distribution about zero, and g(λ) follows the gamma distribution, then we will have,

$$\mu_r^0 = \int_{-\infty}^{\infty} ?(\lambda) m_r d\lambda \qquad (0.0)$$

Where μ_r^0 is the r^{th} moment of the GNB distribution about zero.

We have from Consul and Jain (11) that

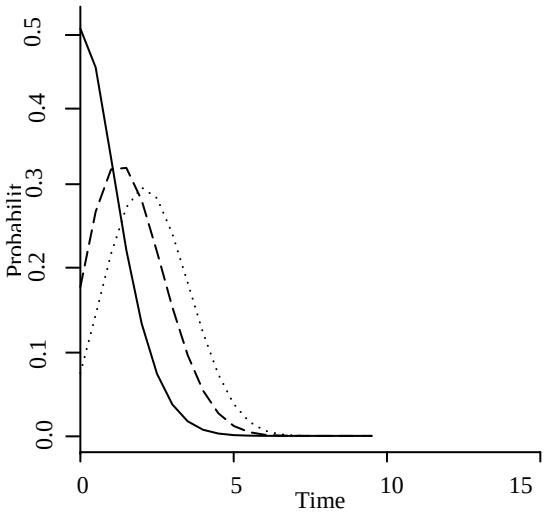

Figure 2: The graphs for fixed values of k = 0.1, β = 3 and with different values of α = 2.5, 6.9 that has been depicted by solid, dashed and dotted lines, respectively

$$m_1 = \frac{\lambda}{1+\lambda k} \qquad (0.0)$$

and

$$m_2 = \frac{\lambda}{(1+\lambda k)^3} + \frac{\lambda^2}{(1+\lambda k)^2} \qquad (0.0)$$

If we put these values instead of m_r and g(λ), then

$$\mu_1^0 = \int_{-\infty}^{\infty} \frac{\lambda}{1+\lambda k} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \qquad (0.0)$$

but

$$\frac{\lambda}{1+\lambda k} = 1 - \lambda k + (\lambda k)^2 (\lambda k)^3 + \cdots \qquad (0.0)$$

Therefore

$$\mu_1^0 = \frac{\alpha}{\beta} - \frac{k\alpha(\alpha+1)}{\beta^2} + \frac{k^2\alpha(\alpha+1)(\alpha+2)}{\beta^3} - \cdots \qquad (0.0)$$

and

$$\mu_{12}^0 = \int_{-\infty}^{\infty} \left[\frac{\lambda}{(1+\lambda k)^3} + \frac{\lambda^2}{(1+\lambda k)^2}\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \qquad (0.0)$$

but

$$(1+\lambda k)^{-2} = \sum_{j=0}^{\infty} \binom{j+1}{j} (-k\lambda)^j \qquad (0.0)$$

and

$$(1+\lambda k)^{-3} = \sum_{j=0}^{\infty} \binom{j+2}{j} (-k\lambda)^j \qquad (0.0)$$

Table 1. Observed and expected number of persons of all ages, by number of doctor's calls or clinic visits in a year

Number of doctor's calls or clinic visits (n)	Observed*	Expected (GNB)	Expected* (NB)
0 calls or visits	7690	7751	7687
1	2044	2021	1986
2	1097	1097	1093
3	718	699	707
4	474	480	492
5	310	344	357
6	250	253	267
7	190	190	202
8	140	145	157
9	111	111	122
10+	514	447	440
Total	13538	13538	13538
χ^2		14.944	25.8
df		7	8
Parameters		$\alpha = 0.317$	$\alpha = 0.307$
		$\beta = 2.208$	$\beta = 0.188$
		$k = 0.006$	

*Taken from Chiang (1965). GNB: Generalized negative binomial, NB: Negative binomial

therefore

$$\mu_2^0 = \left[\frac{\alpha(\alpha+1)}{\beta^2} + \frac{\alpha}{\beta}\right] - \left[\frac{2k\alpha(\alpha+1)(\alpha+2)}{\beta^3} + \frac{3k\alpha(\alpha+1)}{\beta^2}\right] + \cdots \quad (0.0)$$

Thus, mean and variance of the GNB distribution will be smaller than, equal to or greater than the mean and variance of the NB, according to the values of k, if it be positive, zero or negative, respectively.

Application of the GNB distribution: It seems logical that the GNB distribution should give reasonably a good fit for numerical data which follows strictly Poisson and NB or even binomial distribution.

We used some information about the number of illnesses which had occurred in a subpopulation that a NB had been fitted on it by Chiang in 1965. The data in that survey were based on a sample of approximately 10 000 households inflated to give the ratio and figures as appeared in the publication (12). Hence, the published figures are much greater than the actual counts in the sample.

Two indirect measures of illness were used, the number of doctor's calls or clinic visits and the number of complaint periods that individuals have had during the year. For the number of doctor's calls or clinic visits the model was fitted for all ages, and the result was presented in table 1 along with a comparison with the NB. Data, with the number of complaint periods, were used only for the age groups under 15, and the results were shown in table 2.

Table 2. Observed and expected number of persons under 15 years of age, by number of complaint periods in a year

Number of complaint periods (n)	Observed*	Expected (GNB)	Expected (NB)
0 complaint periods	522	545	551
1	875	823	817
2	787	808	799
3	637	652	648
4	458	470	471
5	316	316	318
6	206	201	205
7	125	124	127
8+	190	177	180
Total	4 116	4 116	4 116
χ^2		6.541	6.975
df		5	6
Parameters		$\alpha = 3.44$	$\alpha = 3.113$
		$\beta = 1.25$	$\beta = 1.102$
		$k = -0.006$	

*Taken from Chiang (1965). GNB: Generalized negative binomial, NB: Negative binomial

Comparison of the two sets of expected frequencies for the two distributions (GNB and NB) with the observed one, which has been shown in tables 1 and 2 (13), it is clear that the GNB distribution gives almost a better fit, comparing with the NB distribution.

Possibly the maximum likelihood estimate or another form of estimation method for the GNB distribution throws further light on this subject. These values have been chosen just by the searching methods.

Discussion

In this paper, we extend another formula instead of NB distribution to overcome over-dispersion problem. This distribution is called the GNB distribution with three parameters α, β and k, which is obtained as a compound form of the generalized Poisson and the gamma distributions. Thus, mean and variance of the GNB distribution will be smaller than, equal to or greater than the mean and variance of the NB according to the values of k, if it is positive, zero or negative, respectively.

Two sets of data were analyzed by Chiang (1955, 1965) who used the NB distribution. We proposed a new way called GNB distribution, which was different from the way of Chiang (1965). The result of fitting using GNB along with NB has been shown in tables 1 and 2. The result indicate too much satisfaction. We used a searching method (maximum chi-square) to estimate the parameters. Presumably for the GNB distribution, the maximum likelihood estimates are smaller than the estimates obtained from the estimation method. This paper has shed light on this subject.

Conclusion

NB model has been derived from two different forms. At first, the NB can be thought as a Poisson-gamma mixture. The second form of the NB can be derived as a full member of a single parameter exponential family of distributions, and therefore considered as one of the GLM (generalized linear models). We extend the GNB distribution with three parameter α, β and k, which is obtained as a compound form of generalized Poisson and gamma distributions. This distribution gives a very close fit for a large number of data and provides an appropriate model for numerus studies. With k = 0 the model becomes the ordinary NB and with α = 1 it becomes a new model which we call it the generalized geometric distribution with two parameters. The most important feature of this model is and its time-dependent probabilities. In addition, it can be used for a variety of researches especially in survival analysis with the progressive diseases that the risk of dying changes over time. Hence, this model will be very helpful. We illustrated this model with two data sets, which were indirect measures of illness, by comparing the results of fitting with the ordinary NB. Results were quite satisfactory.

Acknowledgments

Research leading to this paper was supported by the International Campus of Tehran University of Medical Sciences.

References

1. Greenwood M, Yule U. 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. Royal Statistical Society 1920; 83(2): 255-79.
2. Johnson NL, Kotz S. Discrete distributions. New York, NY: Wiley; 1969. p. 328.
3. Takacs L. A generalization of the ballot problem and its application in the theory of queues. J Am Stat Assoc 1962; 57(298): 327-37.
4. Mohanty SG. On a generalised two-coin tossing problem. Biom Z 1966; 8(4): 266-72.
5. Jain GC, Consul PC. A generalized negative binomial distribution. SIAM J Appl Math 1971; 21(4): 501.
6. Cameron AC, Trivedi PK. Regression analysis of count data. 1st ed. Cambridge, UK: Cambridge University Press; 1998. p. 411.
7. Winkelmann R, Zimmermann K. Recent Developments in Count Data Modelling:

Theory and Application. J Econ Surv 1995; 9(1): 1-24.
8. Greene WH. LIMDEP Econometric Modeling Guide. Version 9. Plainview, NY: Econometric Software Inc; 2006.
9. Brass W. The distribution of births in human populations in rural Taiwan. Population Studies 1958; 12(1): 51-72.
10. Dandekar VM. Certain modified forms of binomial and poisson distributions. Sankhya 1955; 15(3): 237-50.
11. Consul PC, Jain GC. A generalization of the poisson distribution. Technometrics 1973; 15(4): 791-9.
12. Canada Dominion Bureau of Statistics, Canada Department of National Health and Welfare. Canadian Sickness Survey 1950-51. No.8. Volume of Health Care (National Estimates). DBS Reference Paper no.51. Ottawa, Canada: The Queen's Printer and Controller of Stationary; 1955.
13. Chiang CL. An index of health: mathematical models. Vital Health Stat 1 1965; (94): 1-19.