International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-3, May, 2014

A Study of Goodness –of- Fit Tests for Some Discrete Probability Distribution

Snigdha Mahanta, Prof. (Dr.) M. Borah

***Abstract*— This paper presents the goodness of fit (GOF) tests for several discrete distributions viz., Poisson, Generalized Poisson and Negative binomial distribution. Parameter estimation is performed and goodness of fit test for a set of real data is obtained for the said distributions.**

***Index Terms*— Poisson distribution, Negative binomial distribution, Goodness of fit, Generalized Poisson distribution, Parameter estimation**

I. INTRODUCTION

Fitting a probability model to observed data is an important statistical problem from both theory and application point of view. Goodness-of-fit techniques have been widely studied and an extensive literature is available. Most work in this field deals with fitting continuous distributions rather than the discrete distributions. Keeping all these points in view, the present author has been attempt to discuss the theory of the particular discrete distribution namely Poisson, Generalized Poisson and Negative binomial distribution and parameter estimation is also performed for this distributions. Finally, the author uses the Chi-square test for testing whether observed data are representative of a particular distribution for a set of real data and conclusions are made.

II. POISSON DISTRIBUTION (PD)

The Poisson distribution is named after Simeon-Denis Poisson (1781–1840). In addition, Poisson is French for fish. A family of probability distributions for a countably infinite sample space, each member of which are called a Poisson distribution. Recall that a binomial distribution is characterized by the values of two parameters: *n* and *p*. A Poisson distribution is simpler in that it has only one parameter. The parameter must be positive. The Poisson distribution is defined mathematically by the formula

$$P(x/\lambda) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, x = 0,1,2,... \\ \end{cases}$$

0, otherwise

Estimation of parameters of Poisson distribution

Simon Denis Poisson (1781-1840) gave the moment estimator for the parameter of PD as:

$$\lambda = \bar{x} = m_1$$

where λ or m_1, is the sample mean.

III. GENERALIZED POISSON DISTRIBUTION (GPD)

The Generalized Poisson Distribution (GPD), introduced in Consul and Jain [1], and studied extensively by Consul [2] is defined on the non-negative integers for $0 \le \lambda_1 \le 1$ and $\lambda_2 > 0$ by

$$Px(\lambda_1, \lambda_2) = \left\{ \frac{\lambda_1(\lambda_1 + x\lambda_2)^{x-1} e^{-\lambda_1 - x\lambda_2}}{x!} \right\}, x = 0,1,2,...$$

0 for x>m, when $\lambda_1 < 0$

Applications of the GPD can be found in settings where one seeks to describe the distribution of an event that occurs rarely in a short period, but where we observe the frequency of its occurrence in longer periods of time. It extends the Poisson distribution by its ability to describe situations where the probability of occurrence of a single event does not remain constant (as in a Poisson process), but is affected by previous occurrences. The distribution has been found (Consul [2], pp. 117–129) to accurately describe phenomena as diverse as the observed number of industrial accidents and injuries, where a learning effect may be present, the spatial distribution of insects, where initial occupation of a spot by a member of the species has an influence on the attractiveness of the spot to other members of the species, and the number of units of different commodities purchased by consumers,

Manuscript received May 10, 2014.
Snigdha Mahanta, Chaiduar College, Sonitpur , Assam, India (e-mail: snigdhamahanta11@gmail.com).
Dr. M. Borah, Professor, Department of Mathematical Science, Tezpur University, Sonitpur, Assam, India (e-mail:mborah@tezu.ernet.in).

where current sales have an impact on the level of subsequent sales through repeat purchases.

The author's defined mean and variance of the GPD as

$$Mean = \frac{\lambda_1}{1-\lambda_2}$$

$$Variance = \frac{\lambda_1}{(1-\lambda_2)^3}$$

And

The variance of this GPD model is greater than, equal or less than the mean according to whether the second parameter λ_2 is positive, zero or negative and both mean and variance tend to increase or decrease in values, as λ_1 increase or decrease.

Estimation of parameters of Generalized Poisson distribution.

Consul and Jain [1] obtained by moment method estimators for the parameter of GPD in the form as:-

$$\lambda_1 = \sqrt{\frac{m_1^3}{m_2}} = \sqrt{\frac{\bar{x}^3}{s^2}}$$

$$\lambda_2 = 1 - \sqrt{\frac{m_1}{m_2}} = 1 - \sqrt{\frac{\bar{x}}{s^2}}$$

Where $\bar{x}$ or m1 and s^2 or m_2 are sample mean and sample variance respectively.

IV. Negative Binomial Distribution (NBD)

The negative binomial distribution (NBD) has appeal in the modelling of many practical applications. A large amount of literature exists, for example, on using the NBD to model: animal populations (see e.g. Anscombe [3], Kendall [4] ;) accident proneness (see e.g. Greenwood and Yule [5], Arbous and Kerrich [6]) and consumer buying behaviour (see e.g. Ehrenberg [7]). Furthermore, the NBD can be implemented as a distribution within stationary processes (see e.g. Anscombe [8], Kendall [9]) thereby increasing the modelling potential of the distribution. The NBD model has been extended to the process setting in Lundberg [10] and Grandell [11] with applications to accident proneness, sickness and insurance in mind.

In case of negative binomial distribution, the following equalities/inequalities are held:

(i) npq > np and $q > \frac{np}{np}$ or q > 1

(ii) since p + q = 1, p must be negative, i.e. p = -p = 1-q

From the above (ii) we have,

-p = 1-q

and q = 1+p , where $p = 1 + \frac{m}{k}$

substituting we get

$$q = 1 + \frac{m}{k}$$

The parameters of the distribution are the arithmetic mean (m) and the exponent k.

Since the variance of the population is,

$$\sigma^2 = kpq = kp(1+p) = kp + kp^2,$$

substituting

$$p = \frac{m}{k}$$ we get,

(iii)

$$\sigma^2 = k\left(\frac{m}{k}\right) + k\left(\frac{m}{k}\right)^2 = m + \frac{m^2}{k}$$

The probability series of the N.B.D. is given by the expansions

$$(q-p)^{-k}$$

The individual terms of $(q-p)^{-k}$ are given by

$$P(x) = q^{-k}\frac{(k+x-1)!}{x!(k-1)!}\left(\frac{m}{m+k}\right)^x$$

By using the recurrence formula the individual terms of the series are,

$$P(x=0) = q^{-k} = \left(1+\frac{m}{k}\right)^{-k}$$ and

$$P(x+1) = \left(\frac{k+x}{x+1}\right)\left(\frac{m}{m+k}\right)P(x)$$

From above (iii) we have,

$$\sigma^2 = kpq = m + \frac{m^2}{k}$$

The above formula indicates that, the reciprocal of the exponent k, i.e., $\frac{1}{k}$ is a measure of the excess of variance or clumping of the individuals in the population. Specifically, as $\frac{1}{k}$ approaches zero and k approaches infinity, the distribution converges to the Poisson series (s2 Þ m). Conversely, if clumping increases $\frac{1}{k}$, 1 approaches infinity (k Þ 0) and the distribution converges to the Logarithmic Series.

Estimation of parameters of Negative Binomial distribution

The estimation of the parameters of NBD by the method of moments given as

$$q = \frac{\hat{\sigma}^2}{\hat{m}}$$ where

$$s^2 = \hat{\sigma}^2 = \frac{1}{\sum_i f_i - 1}\left[\sum_i f_i x_i^2 - \frac{\left(\sum_i f_i x_i\right)^2}{\sum_i f_i}\right]$$

$$\bar{x} = \hat{m} = \frac{1}{\sum f_i}\sum_i f_i x_i$$

-p = 1- q

And

$$\hat{k} = \frac{\hat{m}}{p}$$

V. GOODNESS OF FIT

An attempt has been made to fit the Poisson, Generalized Poisson and Negative binomial distribution by using chi-square test. We have used data sets of road traffic accident under Dibrugarh police station during the period of 2012 for fitting the aforesaid distribution. The expected frequencies according to the methods along with the estimates of the parameters and the values of the chi-square are given in the following table.

Table 1: Data for no. of accident per no. of week of the year 2012

No. of accident	Observed frequency	PD	GPD	NBD
0	16	15	15.3	15.5
1	17	19	18.3	18.2
2	12	12	11.4	11.3
3	4	5	4.9	4.9
4	3	1	1.6	1.6
5	0	0	0.5	0.5
Parameters		$\lambda = 1.25$	$\lambda_1 = 1.22$ $\lambda_2 = 0.020$	$p = 0.063$ $k = 19.94$
$\chi^2 =$		0.44386	0.15595	0.15380
p value		0.8009	0.9249	0.92598

VI. CONCLUSION

What can be whipped away from the analysis made here can be outlined as follows:

From the above table 1, it seems to be very logical that GPD and NBD give comparatively better fit to the data. In the table 1, the estimate of the parameter in Poisson (λ = 1.25), GPD ($\lambda 1$ =1.22, $\lambda 2$ = 0.020) and NBD (p = 0.063, k = 19.94). The computed value of chi-square reflected by their p-value further show that non-significant difference is almost same in GPD and NBD followed by PD. Therefore, it may be conclude that the data fits well for all three distribution viz, Poisson, Generalized Poisson and Negative binomial distribution.

REFERENCES

[1] Consul, P. C. and G. C. Jain., "A generalization of the Poisson distribution.", Technometrics 15(4),1973, 791–799

[2] Consul, P. C., "Generalized Poisson Distributions: Properties and Applications", Volume 99 of Statistics: Textbooks and Monographs. New York: Marcel Dekker Inc., 1989

[3] Anscombe, F. J., "The statistical analysis of insect counts based on the negative binomial distribution.", Biometrics, 5, 165-173. 1949

[4] Kendall, D. G., "On some modes of population growth leading to R. A. Fisher's logarithmic series distribution", Biometrika, 35, 6-15,1948a

[5] Greenwood, M., & Yule, G., "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.", J. Royal Statistical Society, 93, 255-279, 1920

[6] Arbous, A., & Kerrich, J., "Accident statistics and the concept of accident proneness", Biometrics, 7, 340-432, 1951

[7] Ehrenberg, A., "Repeat-buying: Facts, theory and applications", London: Charles Griffin & Company Ltd.; New York: Oxford University Press., 1988

[8] Anscombe, F. J., "Sampling theory of the negative binomial and logarithmic series distributions", Biometrika, 37, 358-382, 1950

[9] Kendall, D. G., "On the generalized "birth-and-death" process.", Ann. Math. Statistics, 19, 1-15., 1948b

[10] Lundberg, O., "On Random Processes and their Application to Sickness and Accident Statistics." , Uppsala: Almquist and Wiksells., 1964

[11] Grandell, J., "Mixed poisson processes", Chapman & Hall., 1997