▼ Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan.
Vol.14, No.2, December 2003, pp. 203-211 ISSN 1021-1012

SOME PROBABILITY DISTRIBUTIONS IN THE CONTEXT OF ROAD ACCIDENTS

G. R. Pasha[1] and Muhammad Akbar Ali Shah[2]
[1]Faculty of Science and Agriculture, Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan.
[2]Department of Statistics, Islamia University, Bahawalpur, Pakistan.

Abstract: In this paper the application of some discrete and continuous distributions in the field of road accidents has been studies. The relationship between these distributions has been explored and their means and variances are estimated. A distribution of vehicles involved in fatal accidents has been generated by Poisson distribution with a mean of 0.5 accidents.

Keywords: Gamma distribution, log series distribution, maximum likelihood, mixed Poisson distribution, multivariate distribution, negative binomial distribution, Poisson distribution, posterior and prior distribution.

RELATIONSHIP BETWEEN DISTRIBUTIONS

The number of accidents at a site is often taken to be Poisson distributed. If the between-site variation is mean accident frequency is taken to be gamma distribution, the resulting overall distribution of accident frequency is negative binomial. The negative binomial can arise as the distributions of the sum of n independent variables each having the same log series distribution where n has a Poisson distribution. The log series distribution can be reviewed as a multivariate distribution consisting as a set of independent Poisson distributions with means αX, $1/2\alpha X^2$, $1/3\alpha X^3$, … If the number of groups of individuals has a Poisson distribution with expected values ϕ and the number of individuals per group has the log series distribution, the distribution of the number of individual is negative binomial with parameters $\alpha\phi$, $\theta/(1-\theta)$.

Gipps [1980] supposed that the number of accidents at a particular site during a fixed period of time was Poisson distribution and that the average number of accidents derived from the Poisson assumption would vary between sites according to a gamma distribution. The combination of these two distributions produced a negative binomial distribution. The parameters for the negative binomial (and the gamma) distribution would be estimated from amassing the data from all sites within the study area.

Abbess *et al*. [1981] also used a negative binomial distribution to test their data. They stated that they obtained satisfactory fits, but because of a problem with too many zero accident sites, they used a truncated negative binomial distribution and obtained good fits except for one of the five years of data. The test used for measuring the goodness of fit was not mentioned.

The logarithmic series distribution derived from Fisher *et al*. [1943]. It was used in a situation where the number of individuals (insects) caught in a

trap was represented by a Poisson variable with expected value λ_i. If the λ's are chosen randomly from a gamma distribution (with origin at zero) then the expected total number of species represented in a given catch by k = 1, 2, 3, … Individuals would be proportional to the terms is a negative binomial distribution, truncated by the exclusion of the first term (k = 0). As the exponent of the negative binomial approaches zero (corresponding to increasing) variability among the λ's) then the probability of taking the value k tends to:

$$\lim_{N\to 0} \frac{\Gamma(N+K)}{\Gamma(N+1)k!} \frac{N}{Q^N - 1} \left(\frac{P}{Q}\right)^k = \alpha \frac{\theta^k}{k}$$

where $\theta = \frac{P}{Q}$

α = -1/ln (1-θ), 0<θ<1 and k = 1, 2, 3, …

Then

$$p(x = k) = \alpha \frac{\alpha^k}{k}$$

The parameter θ may be estimated by reference to tables in Williamson and Bretherton [1964], after calculating the mean number of accidents per intersection. Once θ is determined, the successive values of the probabilities can be found by:

$$P(1) = (1 - \theta)$$

And as $P(n+1) = P(n)\ (n/n+1)^{\theta}\ P(1)\frac{1}{2^{\theta}}$

$$P(2) = P(1)\frac{1}{2^{\theta}}$$

$$P(3) = P(2)\frac{2}{3^{\theta}}$$

$$P(4) = P(3)\frac{3}{4^{\theta}}$$

$$P(5) = P(5)\frac{5}{5^{\theta}}$$

We further suppose that a road network consisting of a number of sites is being studied for two periods of equal length and that the- frequency of accidents at each site is noted for each of the two periods. If proneness is gamma and if, conditional upon a given proneness, accidents follow the Poisson distribution, and if we denote accidents in the first period by the variable, X_1, and accidents in the second period by the variable, X_2, then the bivariate negative binomial given by

$$P(X_1=x_1,X_2=x_2) = \frac{\prod(\alpha + x_1 + x_2)\beta r^{\alpha x_2}(\beta + r + 1)^{-(\alpha+x_1+x_2)}}{\prod(\alpha)x_i!\,x_2!}$$

may be used to give the probability of a given pair of observed frequencies of accidents for the two periods. This is a special case of the distribution considered by Bates and Neyman [1952]. The means of X_1 and X_2 respectively are given by

$$E(x_1) = \alpha/\beta \text{ and } E(X_2) = \alpha r/\beta$$

These expressions represent the mean number of accidents per site in periods one and two respectively. The parameter r may be used to represent the effects of trends or treatment on accident rates. Clearly if it is equal to one then the mean accident rate in the two periods is the same. If r is less than one the accident rate is the second period is reduced and if it is greater than one the rate is increased.

During the first period of study the sites with the highest observed accident rates, say with some rate $x_1 > k$, have been chosen to have remedial treatment. For the second period of study we will therefore have two classes of site: untreated and treated. It is supposed that for sites in the first category $r = r_1$ whereas for the sites in the second category $r = r_2$. If without loss of generality, we number the sites so that the first m_1 are untreated and the next m_2 are treated, there being n sites in total and we use a further subscript, i to denote site, then the likelihood function for the data for the n sites for the two periods may be written:

$$L(\alpha,\beta,r_1,r_2)=\prod_{i=1}^{n}\frac{\prod(\alpha + x_{1i} + x_{2i})\beta r^{\alpha x_{2i}}(\beta + r + 1)^{-(\alpha+x_{1i}x_{2i})}}{\prod(\alpha)x_{1i}!\,x_{2i}!}$$

Where $r = r_1$, if $x_1 < k$ and $r = r_2$, if $x_1, \geq k$.

The parameter r_1 represents the proportion by which the mean accident rate is affected due to the operation of any secular trends and the ratio of r_2, to r_1 represents the differential proportionate effect due to treatment of the accident black spots. We define S_2 as the total number of accidents for the first period for the untreated Group and T_1 as the corresponding total for the group to be treated. Similarly S_2 is the total of accidents in the second period for untreated sites and T_2 is the total number of accidents for the second period for the treated group. We let $S_2 = S_1 + S_2$ be the total number of accidents for the untreated group over the two periods and $T_3 = T_1 + T_2$ be the corresponding total for the treated group. We further let $y_1 = x_{1i} + x_{2i}$ be the total for the site I for the two periods. Maximizing the likelihood yields the following equations

$$\gamma_1 = S_2(\beta + 1)/(m_1\alpha + S_1)$$
$$\gamma_2 = T_2(\beta + 1)/(m_2\alpha + T_1)$$
$$\beta = n\alpha/(S_1 + T_1)$$

$$\sum_{i=1}^{n} [1/\alpha+\ldots+1]/(\alpha+y_1-1)+n \log [n\alpha/(n\alpha+S_1+T_1)]$$

$$- m_1 \log [(m_1\alpha+S_3)/(m_1\alpha+S_1)] - m_2 \log [(m_2\alpha+T_3)/(m_2\alpha+T_1)][=0$$

$$\alpha = \bar{x}_1^2 /(\sigma_1^2 - \bar{x}_1)$$

where $\bar{x}$ is the mean accident rate for all sites in the first period and σ^2 is the variance.

APPLICATION OF GAMMA DISTRIBUTION

We assume that at any particular black spot in the absence of treatment, accidents occur in a Poisson process of constant true rate m per year. Thus if a denotes the number of accidents at the site in a particular year, a has a Poisson distribution P(a/m) with mean m so that

$$P(a/m) = \frac{\bar{e}^m m^a}{a!} \qquad a = 0,1,2,\ldots$$

Moreover, m is constant over time and the number of accident in different years are independent random variables each having the Poisson distribution.

The true accident rate m will vary from site to site, and its value for any particular site is unknown, but we will regard this value as a random variable. We suppose that the prior distribution of m is described by a probability density function f_0 (m). It is mathematically convenient to assume that this prior distribution is a gamma distribution with parameters n_0 and S_0. So that

$$f_0(m)\frac{n_0(n_0m)^{S_{0-1}}e^{-n_om}}{\Gamma(S_0)} \qquad m > 0$$

Gamma distribution p.d.f. is

$$f_0(m) = \frac{n_0(n_0m)^{S_{0-1}}e^{-n_om}}{\Gamma(S_0)} \qquad m > 0$$

Writing as follow

$$f_0(m) = \frac{n_0^{S_0} m^{S_0-1} e^{-m/(\frac{1}{n_0})}}{\Gamma(S_0)(\frac{1}{n_0})^{S_0}} \qquad m > 0$$

By definition

$$\text{Mean} = E(m) = \int mf(m)dm$$

$$= \int_0^\infty \frac{m m^{S_{0-1}}}{\Gamma S_0 (\frac{1}{n_0})^{S_0}} e^{-m/(1/n_o)} dm$$

$$= \frac{1}{\Gamma S_0 (\frac{1}{n_0})^{S_0}} \int_0^{\infty} m^{S_{0+1-1}} e^{-m/(\frac{1}{n_0})} dm$$

Comparing with Gamma function with two parameters

$$\int_0^{\infty} X^{\alpha-1} e^{-x/\beta} dX = \Gamma \alpha \beta^{\alpha}$$

$$\int_{0\infty}^{\infty} m^{S_{0+1-1}} e^{-m/(\frac{1}{n_0})} dm = \Gamma S_0 + 1 (\frac{1}{n_0})^{S_{0+n}}$$

$$E(m) = \frac{1}{\Gamma S_0 (\frac{1}{n_0})^{S_0}} \Gamma S_0 + 1 (\frac{1}{n_0})^{S_{0+1}}$$

$$= \frac{S_0 \Gamma S_0 (\frac{1}{n_0})^{S_0} (\frac{1}{n_0})}{\Gamma S_0 (1/n_0)^{S_0}} \qquad = s_0 / n_0$$

$E(m)$ = s_o/n_o = Mean variance

$E(m^2)$ = $\int_{0\infty}^{\infty} m^2 f(m) dm$

$$= \int_{\infty}^{\infty} \frac{m^2 m^{S_0 - 1} e^{-m(1/n_o)} dm}{\Gamma S_0 (1/no)^{S_0}}$$

$$E(m^2) = \int_{\infty}^{\infty} \frac{m^{S_0+2-2} e^{-m(1/n_0)} dm}{\Gamma S_0 (1/no)^{So}}$$

Comparing with the Gamma function

$$E(m^2) = \frac{\Gamma So + 2 (1/no)^{S_{0+2}}}{\Gamma So (1/no)^{So}} = \frac{So(So+1) \Gamma S_0 . (1/n_o)^2 (1/n_o)^{So}}{\Gamma S_0 (1/n_o)^{So}}$$

$$E(m^2) = \frac{So(So+1)}{n_0^2}$$

For variance

$$V(m) = E(m^2) - E(m)^2 = \frac{S_o(S_o+1)}{n_0^2} \left(\frac{S_o}{n_0} \right)^2 \quad V(m) = S_o / n_o^2$$

This distribution has mean S_o / n_o and variance S_o / n_o^2. If the prior distribution is of gamma type with parameters S_0 and n_0, and if in a period of n years the observed total number of accidents is S_1, then, under the

assumptions of Poisson and gamma distribution, the posterior distribution of m is also of gamma type, but with parameters S_1 and n_1, Where

$$S_1 = S_0 + S$$
$$n_1 = n_0 + n$$

The prior distribution contains an amount of information about the value of m, which is equivalent to the amount we would gain if So accidents had actually occurred at the site over a period of n_0 years [Raiffa and Schlaifer 1961, Maritz 1970]. If the posterior distribution is of the gamma type with parameters n_1 and S_1 than it is known that this predictive distribution will be of the negative binomial form given by:

$$P(a)\frac{\Gamma(S+a)}{a!\,\Gamma(S_1)}\left(\frac{1}{1+n_1}\right)^{a}\left(\frac{n_1}{1+n_1}\right)^{S_1} \qquad a = 0,1,2.....$$

The distribution is more dispersed than the Poisson its variance is greater than its mean, whereas the mean and the variance of the Poisson distribution are equal. This greater dispersion arises from the fact that our knowledge of the true rate m is imprecise. If we know the true value precisely we could say straight away that P(a) would be Poisson with mean m. The above distribution describes the random variation in accident from year to year, after taking into account our uncertainty about the mean itself. It can be shown that as the amount of accident in formation, incorporated in the parameters n_1 and S_1 increases, P(a) does in fact approximate more and more closely to a Poisson distribution.

If we assume that the number of accidents at the various sites is independent random variables, then we will have

$$q(a) = \int_0^{\infty} P(a/m) f_0(m)\,dm$$

Now if $f_o(m)$ is of the gamma type, then it can be shown that q(a) will be negative binomial. The converse is also true if q(a) is negative binomial. Then the distribution of m must be gamma [Maritz 1970].

Table 1: Accident involved vehicles generated by Poisson distribution with a mean of 0.5 accidents in 1988.

Number of fatal accidents	0	1	2	3	4	5	6	7	8
Number of vehicles	467	234	58	10	1	0	0	0	0

If we take a figure of 0.5 accidents in a study year as the expected mean, Table1 shows the distribution of the number of actual accidents one would expect to get among a sample of 770 identical vehicles in 1988 in Karachi if the accidents occurred strictly by chance. The mean is obviously 0.5 and the variance (defined algebraically as $\Sigma(x - m)^2/n$, where x is an observation i.e. 0,1,2….8, m is the mean and n the number of observations) is 0.5 also. It is a property of the Poisson distribution that the variance equals the mean. Seeing table, the question naturally arises

as to whether those vehicles involved in two or more fatal accidents have a higher accidents liability than the rest.

In Table 1, since we defined the set of 770 vehicles as “identical” we can say that they do not. In the next year they would have exactly the same chance as everyone else of having 0,1,2 or 3 fatal accidents, and there is a very highly probability that they would involved fever accidents, the regression to the mean effect [Abess *et al*. 1981, Hauer 1980, Wyshak 1974, Campbell 1974, Gipps 1980].

In reality, of course, no matter how a sample of vehicles involved in accident is chosen, there are going to be differences between vehicles in accident liability. Some factors like distances travelled per year, we would expect to vary from vehicle to vehicle with consequent variation in the expected number of accidents. Other factors, age life, vehicle design, maintenance of the vehicle may well influence accident liability also; for the moment we are not concerned with the reasons for the variability, but only with the statistical consequences. Let us assume, that, unlike the previous group of ‘Identical’ vehicles, we choose a new group of vehicles who have been mean accident frequencies per year that vary from 0 to 1,0; the distribution of means is rectangular so that there is an equal probability of any driver having a mean accident frequency within the above range. The actual number of accidents in which vehicles are involved will have of course still be 0, 1, 2, or 3 but the distribution will not be the Poisson distribution. It will be mixed Poisson with a variance, which is greater than that of the Poisson. If accidents still happens randomly to our new set of vehicles, but at rates determined by the new distribution of means then the resulting distribution of accidents will have a mean of 0.5 as before, but a variance of about 0.58- that is a variance which is greater than the Poisson variance by an amount (0.08) which arisen from the variation between means.

As a matter of fact, it is much simpler algebraically if some assume that the underlying distribution of means is not rectangular, but of Gamma form. In this case the sampling distribution of accidents in a given period of time is exactly calculate as a negative binomial. The variances of these three distributions given that the mean value for each is, say, m, are as follows;

Poisson	:	m
Gamma	:	m^2/S
Negative Binomial	:	$m + m^2/s$

S is the 'Parameter' of the gamma distribution; it effectively changes the 'spread' of the distribution so that if S is small (1, say) the gamma distribution is broad, whereas as S gets bigger the distribution becomes moreover. For the gamma distribution the' coefficient of variation is $1/\sqrt{3}$ independent of the mean -that is to say that this particular model of accidents implies that the spread of the underlying accident liabilities is

proportional to the mean value or, but another way, is constant in percentage terms.

CONCLUDING REMARKS

The important point is the order of magnitude of the components of variance. If m = 0.5 as before, and S is about 6, then 2/S = 0.04. That is to say, if we are interested in studying the underlying variations in the mean values of accident liability and the random element is simply a 'nuisance factor' then we are interested missing in the data. If m is higher (more spread in distribution of underlying means), then the proportion of variability in the observations, which is of interest to us increases and vice versa. In these circumstances, the only way to detect the relatively small variations of interest against a high background of unwanted variably is to use large samples.

The other factors of this situation, which has given rise to some frustration among researchers is the difficulty of obtaining any apparently meaningful levels of correlation between accidents and other measured parameters, measures of vehicle maintenance. Even correlations between accidents in one period with those in a subsequent period (which ought to exist if there are real and permanent differences of accident liability between vehicles) typically yield correlation coefficients of less than 0.3. But again this is a direct consequence of the fact that the variability of interest is swamped by random variation in the observations. It has been shown that if the underlying distribution of means is of Gamma form, then the expected number of accidents in a future period, N for a group of vehicles involved in N_C accidents in a current period (assuming for simplicity the two periods are of equal duration) is related as follows:

$$N_t \text{ (expected)} = K (1 = N_C / S)$$

Where K= m 1(1+ m/S) and m and S have the some meaning as before. The variance of N_t also increases linearly with N_C but is always greater than the basic Poisson variance. So attempts to correlate accidents in a future period with those in a current period will yield low correlation coefficients, even if there is a real underlying effect of practical significance [Sabey and Staughton 1975, Johnson and Garwood 1957].

References

Abess, *et al*. (**1981**) “Accidents at blackspots estimating the effectiveness of remedial treatment with special reference to the regression to mean effect”, *Traffic Engineering and Control*, 22(10), 535-542.

Campbell, B.S. (**1974**) “Objective programme evaluation”, *Proceedings of 7th Australian Road Research Board Annual Conference*, Adelaide, Paper No. S4/A44.

Gipps, P.G. (**1980**) “Examining the safety contributions of traffic control devices”, *World Conference On Urban Transport Research*, London. Paper No. E26.

Hauer, E. (**1980**) “Selection for treatment as a source of bias in before-and-after studies”, *Traffic Engg. and Control*, 8/9, 419-421.

Johnson, N.I. and Garwood, F. (**1957**) “Analysis of the claim records of a motor insurance company”, *Journal of Actuaries*, 83 (Part III), 365.

Maritz, J.S. (**1970**) “Empirical Bayes Methods”, Methuen.

Raiffa, H. and Schlaifer, R. (**1961**) “Applied Statistical Decision Theory”, Harward Business School, pp. 283-289.

Sabey, B.E. and Staughton (**1975**) “Interacting roles of road environment vehicle and road user in accidents”, *Paper presented in the 5th International Conference of the Association for Accidents and Traffic Medicine*, London.

Szulga, J. (**1998**) “Introduction to Random Chaos”, Chapman and Hall, CRC, UK.

Williamson, E. and Bretherton, M. (**1964**) “Tables of the logarithmic series distribution”, *Annals of Mathematics and Statistics*, 35, 284-297.

Wyshak, G.A. (**1974**) “A programme for estimating the parameters of the truncated negative binomial distribution”, Algorithm A568, *Applied Statistics*, 23(1), 87-91.