

UNIQUE PROPERTIES OF SOME DISTRIBUTIONS AND THEIR APPLICATIONS

John Panaretos, University of Missouri - Columbia

ABSTRACT

In many practical situations bivariate probability distributions are used whose marginals are of the same form. Sometimes however, in cases of a not too good fit, one of the marginals appears to describe the corresponding observed data exceptionally well while the other provides a rather poor fit. The bivariate model then has to be questioned. This paper suggests ways in which characterization theorems can be used to explain this paradox and also guide the investigator's choice towards possible alternative models that might provide a better fit.

Key words and Phrases: Poisson Distribution, Negative Binomial Distribution, Binomial Distribution, Compounding, Characterization, Accident Statistics.

1. Introduction.

Bivariate probability models are very frequently used to describe random phenomena, mainly because they provide a deeper insight into the underlying chance mechanism. In many of the cases where a bivariate model is used the conditional distribution of one of the random variables given their sum is very informative. In the discrete case these models involve the non-negative, integer-valued random variables X, Y, Z where $X = Y + Z$ and also the conditional distribution of Y on X . A special class of these models includes a situation where X and Y have the same distribution. An interesting example of a simple model of this nature is the one adopted by Leiter and Hamdan (L & H) in 1973, in their study of traffic accidents. In fact they used the bivariate model (X, Y) to express the relationship between the number X of accidents at a certain location during a given time interval and the number Y of fatal accidents among the X accidents. They assumed X to follow a Poisson distribution with parameter λ . Then, by letting p be the probability that an accident is fatal and by assuming that accidents occur independently of one another and that p is constant, they found that Y is also Poisson (λp). So they ended up with the bivariate model

$$f(x,y) = e^{-\lambda} \lambda^x p^y q^{x-y} / y! (x-y)! \quad (1.1)$$

$$x, y = 0, 1, 2, \dots; y \leq x;$$

$$0 < p < 1, q = 1 - p; \lambda > 0.$$

In applying this model to some accident data they found that the fit was not very satisfactory. It is interesting however, that the fit of the Poisson to the distribution of the number Y of fatal accidents

was very good while the fit of the Poisson to the distribution of the number X of accidents was rather poor. This peculiarity, which also occurs in other problems of similar nature, is part of what this paper deals with. Specifically, our aim is to examine whether there is any theoretical justification or explanation of why this happens. We also try to see how theoretical knowledge can help in constructing alternative models in such cases. For this, we appeal to some characterization theorems for discrete distributions.

In a recent paper Kekalaki and Panaretos (1979) studied some characterizations of the Poisson and the compound Poisson distributions. The potential use of these characterizations in problems like the one we are dealing with was mentioned. So, the present paper is a continuation of that work in the sense that it actually utilizes the theoretical results in this specific practical problem.

In the sequel, we state the characterizations mentioned above. We then concentrate our attention to the results related to the L & H model and point out how these characterizations can guide the investigator to models that do not have the deficiencies mentioned earlier. Finally, we actually apply some of the models to the data of L & H and discuss the improvement achieved.

2. The Theoretical Results.

Theorem 1. (Kekalaki and Panaretos (1979)).

Suppose that for the non-negative, integer-valued random variables X, Y considered in the introduction we have that

$$P(Y = y | X = x) = \int_0^1 \binom{x}{y} p^y q^{x-y} dF(p);$$

$$y = 0, 1, \dots, x; x = 0, 1, \dots$$

i.e., that $Y | (X = x)$ is binomial compounded on p by a distribution with distribution function $F(p)$. ($Y | (X = x) \sim b(x, p) \wedge F(p)$). Then Y is Poisson (λp) $\wedge F(p)$ if and only if (iff) X is Poisson (λ).

Theorem 2. (Kekalaki and Panaretos (1979)).

Let X, Y be random variables as in Theorem 1. Assume that the distribution of X is determined uniquely by its factorial moments and that

$$\int_0^\infty \lambda^x dG(\lambda) < \infty \text{ for } \lambda > 0, \quad x = 0, 1, \dots$$

Then

$$Y \text{ is Poisson } (\lambda p) \wedge G(\lambda) \wedge F(p)$$

$$\text{iff } X \text{ is Poisson } (\lambda) \wedge G(\lambda).$$

Of these two general results we will

concentrate our attention on three special cases which are of interest in our study. First, from Theorem 1 we can see that if $F(p)$ is degenerate, then, on the assumption that $Y|(X=x)$ is $b(x,p)$, X is Poisson (λ) iff Y is Poisson (λp). So, in our accident problem under L & H 's assumptions we can be sure that if one of the marginals is Poisson the other has to be Poisson. The fact that in L & H 's study the Poisson fits well one of the marginals while this is not so with the other, implies that the entire Poisson model has to be questioned or that the assumption that p is constant has to be dropped. Because of the characterization there is no possibility that a distribution other than the Poisson can be fitted to X while Y is Poisson and $Y|(X=x)$ is binomial. Further, once one of the marginals is found not to give a good fit the bivariate model (1.1) cannot be expected to be very efficient.

Theorems 1 and 2 suggest two other directions in which one can pursue this problem in order that both marginals be of the same form. One is to allow the parameter λ of the Poisson distribution to be a random variable (indicating a difference in the accident rate from individual to individual). The distribution most commonly used for λ in such cases is the gamma (k,m), i.e.,

$$dG(\lambda) = \frac{m^{-k}}{\Gamma(k)} e^{-\lambda} \lambda^{k-1} d\lambda \quad \lambda, k, m > 0$$

where $\Gamma(a)$ is the usual gamma function.

Another alternative is to drop the assumption that p (the probability of an accident being fatal) is constant from accident to accident. Instead, we can assume that p is a random variable following a beta distribution of type I ($p \sim \text{beta I}(a,b)$). i.e.,

$$dF(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp$$

$$0 < p < 1, \quad a, b > 0.$$

In what follows these two models are examined in detail.

3. The Negative Binomial-Binomial Model (NB-B Model).

As a corollary of Theorem 2 we can easily see that if $F(p)$ is degenerate and $G(\lambda)$ is gamma (k,m) then $X \sim \text{Poisson}(\lambda) \wedge \text{gamma}(k,m)$ iff $Y \sim \text{Poisson}(\lambda p) \wedge \text{gamma}(k,m)$. This is equivalent to saying that $X \sim \text{negative binomial}(k, \frac{1}{m+1})$ iff $Y \sim \text{negative binomial}(k, \frac{1}{1+mp})$. (i.e., $X \sim$

$\text{NB}(k, \frac{1}{m+1})$ iff $Y \sim \text{NB}(k, \frac{1}{1+mp})$. We call this the NB-B model).

This characterization implies that for constant p , if the NB fits well the distribution of Y it should also fit well the

distribution of X , otherwise the whole model will be rather inappropriate. Moreover, under this model if the fit for both X and Y is good then the fit for (X,Y) with

$$f(x,y) = \frac{k(x)}{y!(x-y)!} \left(\frac{mq}{m+1}\right)^x \left(\frac{1}{m+1}\right)^k \left(\frac{p}{q}\right)^y$$

$$q = 1 - p, \quad k(x) = k(k+1)\dots(k+x-1), \\ \cdot k(0) \equiv 1.$$

is expected to be satisfactory.

To apply this model to the data of L & H the method of moments has been used. The parameters of interest are m, p, k . From the above we can easily find that $E(X) = km$, $E(Y) = kmp$ and $V(X) = km(1+m)$.

Hence, the moment estimators of m, p and k are

$$\hat{m} = \frac{S_X^2 - \bar{X}}{\bar{X}}, \quad \hat{p} = \frac{\bar{Y}}{\bar{X}}, \quad \hat{k} = \frac{\bar{X}^2}{S_X^2 - \bar{X}}$$

where \bar{X}, \bar{Y}, S_X^2 are the moment estimators of $E(X)$, $E(Y)$ and $V(X)$ respectively.

4. The Negative Binomial, Negative Hypergeometric Model (NB-NH Model).

As it was mentioned in Section 2 another alternative to the Poisson-binomial model of L & H is to let p be a random variable following a beta I distribution. In this case $Y|(X=x) \sim b(x,p) \wedge \text{beta I}(a,b)$, i.e., negative hypergeometric $h(x;a,b)$. Then from Theorem 1 we have that if $\lambda \sim \text{gamma}(h+l, m)$ then $X \sim \text{Poisson}(\lambda) \wedge \text{gamma}(h+l, m)$ iff

$$Y \sim \text{Poisson}(\lambda p) \wedge \text{gamma}(h+l, m) \wedge \text{beta I}(h, l).$$

It can easily be seen that this is equivalent to

$$X \sim \text{NB}(h+l, \frac{1}{m+1}) \text{ iff } Y \sim \text{NB}(h, \frac{1}{m+1}).$$

(We call this the NB-NH model.)

The reader may observe that the parameters of the gamma and the beta I distribution are suitably chosen so that the distributions of X and Y are of the same form, a property desired for our model. Comments similar to those made in the previous section connecting this characterization to the actual fit of the model to L & H 's accident data can also be made here. Of course, if this model is good then the only possible form for the distribution of (X,Y) will be

$$f(x,y) = \frac{h(y)}{y!} \frac{l(x-y)}{(x-y)!} \left(\frac{m}{m+1}\right)^x \left(\frac{1}{m+1}\right)^{l+h}$$

It can be easily seen that with the above model

$$E(X) = (h+l)m, \quad E(Y) = hm, \quad V(X) = m(m+1)(h+l).$$

The parameters of interest are m, h, l and

their moment estimators are

$$\hat{m} = \frac{S_X^2 - \bar{X}}{\bar{X}}, \hat{h} = \frac{\bar{Y}}{\hat{m}}, \hat{l} = \frac{\bar{X} - \bar{Y}}{\hat{m}}$$

5. Fitting the Models to L & H's Data.

We have fitted the models discussed in Sections 3 and 4 to the accident data used by L & H. These were data of accidents and fatal accidents which occurred in a 50-mile stretch of Interstate 95 in Prince Williams, Stafford and Spottsylvania counties in eastern Virginia. They were collected by the Virginia State Police from 1 January 1969 to 31 October 1970. The two models were fitted to the observed joint distribution of the number of injury accidents and the number of corresponding fatal accidents for the 639 days and for each of the individual years. The fit of each model is measured by the chi-square goodness of fit criterion. The first entry in each cell of the tables is the actual observation. The second entry is the expected frequency under the NB-B model of Section 3. The third entry represents the expected frequency with the NB-NH model of Section 4. Finally, the fourth entry is the expected frequency corresponding to the P-B model studied by L & H. From the tables it becomes clear that the NB-B and the NB-NH provide a better fit as compared to that of the P-B model of L & H. Of course this was somewhat expected, since in the new models more parameters are involved. However, the interesting thing coming out of this study is perhaps the fact that a suggestion was given as to how finding unique properties of distributions can help in guiding the investigator's choice of better models.

It should also be pointed out that the models of Sections 3 and 4 were chosen among the different models offered by Theorems 1 and 2 mainly because they provide the same form for the marginal distributions of X and Y and they were used merely as examples to illustrate the main point. There may be other models that also describe well the problem in question.

Remark: It is worth pointing out that for our models the assumption made in Theorem 2 that the distribution of X is determined uniquely by its factorial moments is redundant. This is so, because the negative binomial distribution is indeed uniquely determined by its factorial moments.

Table 1. Observed and fitted distributions for the number of injury accidents and the number of fatal accidents for the entire study. (First entry: observed frequencies. Second entry: estimated NB-B frequencies. Third entry: estimated NB-NH frequencies. Fourth entry: estimated P-B frequencies (L&H).)

		Number of Fatal Accidents (Y)		
		0	1	TOTAL
0		286		286
		285.25		285.25
		285.25		285.25
		269.78		269.78
1		198	18	216
		201.82	13.69	215.51
		201.82	13.69	215.51
		217.85	14.78	232.63
2		82	10	92
		83.1	11.27	94.37
		83.89	9.69	93.58
		87.96	12.34	100.30
3		24	6	30
		26.02	5.30	31.32
		26.71	4.03	30.74
		23.68	5.15	28.83
4		13	1	14
		6.87	1.86	8.73
		7.21	1.28	8.49
		4.78	1.43	6.21
5		1	0	1
		1.61	0.54	2.15
		1.73	0.35	2.08
		0.89	0.36	1.25
TOTAL		604	35	639
		604.67	32.66	637.33
		606.61	29.04	635.65
		604.96	34.06	639.00

Estimates of the parameters:

NB-B model: $\hat{p} = .0635209$
 $\hat{m} = .1413161, \hat{k} = 6.1018133$

NB-NH model: $\hat{m} = .1413161$
 $\hat{h} = .3875925, \hat{l} = 5.7142208$

VALUE OF THE CHI-SQUARE STATISTIC

Model	χ^2	Degrees of Freedom (v)	$P(\chi_v^2 \geq \chi^2)$
NB-B	8.4969	7	0.30
NB-NH	8.1064	7	0.34*
P-B (L&H)	19.1187	8*	0.02*

*Leiter and Hamdan (1973) used 10 degrees of freedom instead of the correct 8. As a result the p-value of their test was given as 0.04.

Table 2. Observed and fitted distributions for the number of injury accidents and the number of fatal accidents for 1969. (Entries as in table 1).

		Number of Fatal Accidents (Y)		
		0	1	TOTAL
Number of Injury Accidents (X)	0	154		154
		153.94		153.94
		153.94		153.94
		144.81		144.81
1	107	12	119	
	109.03	8.42	117.45	
	109.03	8.42	117.45	
	118.25	9.13	127.38	
2	43	6	49	
	45.33	7.00	52.33	
	45.84	5.96	51.80	
	48.28	7.74	56.02	
3	15	4	19	
	14.42	3.34	17.76	
	14.88	2.51	17.39	
	13.14	3.29	16.93	
4	7	0	7	
	3.89	1.20	5.09	
	4.12	0.81	4.93	
	2.68	0.93	3.61	
5	1	0	1	
	0.93	0.36	1.29	
	1.02	0.22	1.24	
	0.51	0.24	0.75	
TOTAL	327	22	349	
	327.54	20.32	347.86	
	328.83	17.92	346.75	
	327.67	21.33	349.0	

Estimates of the parameters:
 NB-B model: $\hat{p} = 0.0716612$,
 $\hat{m} = 0.1529634$, $\hat{k} = 5.7507616$
 NB-NH model: $\hat{m} = 0.1529634$,
 $\hat{h} = 0.4121067$, $\hat{l} = 5.3386549$

VALUE OF THE CHI-SQUARE STATISTIC

Model	χ^2	Degrees of Freedom (v)	$P(\chi^2_{\nu} \geq \chi^2)$
NB-B	6.0390	7	0.54
NB-NH	5.6931	7	0.58
P-B (L&H)	12.5399	8*	0.14*

*See comment at the end of table 1
 (L&H: 10 d.f.; $P(\chi^2_{10} \geq 12.5399) \approx 0.25$).

Table 3. Observed and fitted distributions for the number of injury accidents and the number of fatal accidents for 1970 (Entries as in table 1).

		Number of Fatal Accidents (Y)		
		0	1	TOTAL
Number of Injury Accidents (X)	0	132		132
		131.47		131.47
		131.47		131.47
		125.02		125.02
1	91	6	97	
	92.70	5.22	97.92	
	92.70	5.22	97.92	
	99.59	5.61	105.20	
2	39	4	43	
	37.72	4.24	41.96	
	38.00	3.68	41.68	
	39.66	4.59	44.25	
3	9	2	11	
	11.60	1.96	13.56	
	11.84	1.51	13.35	
	10.53	1.88	12.41	
4	6	1	7	
	2.99	0.67	3.66	
	3.11	0.47	3.58	
	2.48	0.64	3.12	
TOTAL	277	13	290	
	276.48	12.09	288.57	
	277.12	10.88	288.00	
	277.28	12.72	290	

Estimates of the parameters:
 NB-B Model: $\hat{p} = 0.0532787$,
 $m = 0.1297294$, $k = 6.4856475$
 NB-NH model: $\hat{m} = 0.1297294$,
 $h = 0.3455468$, $l = 6.1401007$

VALUE OF THE CHI-SQUARE STATISTIC

Model	χ^2	Degrees of Freedom (v)	$P(\chi^2_{\nu} \geq \chi^2)$
NB-B	3.9800	5	0.56
NB-NH	4.3395	5	0.50
P-B (L&H)	6.6696	6*	0.37*

*See Comments at the end of table 1.
 (L&H: 8 d.f.; $P(\chi^2_8 \geq 6.6696) \approx 0.55$).

REFERENCES

- Leiter, R.E. and Hamdan, M.A. (1973).
 Some Bivariate Probability Models
 Applicable to Traffic Accidents and
 Fatalities. Int. Stat. Rev. 41(1),
 87-100.
 Kekalaki, Evdokia and Panaretos, J. (1979).
 Characterizations of Compound Poisson
 Distributions. Int. Stat. Inst. Bull.,
 48, 577-580.