Journal of Hydroscience and Hydraulic Engineering
Vol.12, No.2, November, 1994, 113-121

ANISOTROPIC BI-VARIATE DISCRETE STOCHASTIC MODEL FOR HYDROLOGICAL SERIES IN THE DRY SEASON

BY

Masashi Nagao
Department of Civil Engineering, Nagoya Institute of Technology, Nagoya, Japan

Hirokazu Konishi
Research Institute, Water Resources Development Public Corporation, Saitama, Japan

and

Masato Suzuki
Department of Civil Engineering, Gifu National College of Technology, Gifu, Japan

SYNOPSIS

This study presents an extension of the discrete stochastic model used to deal with anisotropic bi-variate hydrological quantities such as precipitation or runoff discharge in the dry season. We illustrate the theory of discrete distribution as the forms of bi-variate binomial distribution or bi-variate negative binomial distribution. In these two types of distributions, each bi-variate has some population parameters such as upper boundary, shape parameter and mutual correlation. By separating marginal and joint distribution we discussed the estimation of parameters in a moment method and a maximum likelihood method. We applied the negative binomial model to the data series of runoff discharge in the dry season. The results prove that the data fit well for the model.

INTRODUCTION

In the case of the study in suitable judgment and optimal operation of utilities on water-resources system, a mathematical model plays an important role in the stochastic distribution of hydrological amounts, for example stream-flow or precipitation. Because of computational necessity, hydrologic engineers have often used the theory of discrete distribution. The authors reported the applicability of the discrete isotropic bi-variate distribution model for reasonable water usage in a reservoir system (1). Recently, we presented the theory on an anisotropic bi-variate model. This study aims at the development of the fundamental statistical characters and the practical parameter estimation for anisotropic discrete bi-variate binomial or negative binomial distribution. We can show the availability of this distribution, by considering the mutual correlation of the runoff discharge series of two adjacent catchment areas in the dry season.

The following stochastic characters of probability density are essential for a stochastic model building in the arid season.

a) flexibility for various positive skewness in accordance with runoff condition
b) dominant persistence in the time-series
c) importance of quantitative representation for the smaller domain of hydrological quantity

Under the computational condition in the relation to the above c), various attempts to introduce the approximate upper boundary have been carried out. We look at the bi-variate discrete distribution theory in this paper.

ANISOTROPIC BI-VARIATE DISCRETE DISTRIBUTION MODEL

The authors reported the deduction of theory and application of a discrete distribution model (2). One is a bi-variate binomial discrete distribution that has common upper boundary (3). Another is an anisotropic bi-variate negative binomial model that has no upper boundary but has common population parameters (4).

The following relation, mean>variance, holds for binomial distribution, but, mean<variance, holds for negative binomial distribution. Because we do not have enough space to describe details about the deduction of this model, we will introduce an outline of those anisotropic bi-variate models.

Binomial distribution Model

We believe that this model will be useful for statistical hydrology. A typical example is the runoff series from two adjacent catchment areas. In the model we assume that the variables have the similar upper boundaries on rainfall intensity or specific discharge.

1) Conditional distribution

Let's assume that two variables X_1 and X_2 have a common upper boundary r but different shape parameters. The probability generating function (p.g.f.) on X_1 and X_2 is given as follows (5):

$$G_{X_1,X_2}(z_1,z_2)=[A+B_1z_1+B_2z_2+Cz_1z_2]^r, \quad A+B_1+B_2+C=1 \tag{1}$$

From this, we get the conditional distribution of $X_1=j$ for a fixed value $X_2=i$ with the shape parameters a_1, a_2 and a mutual correlation ρ,

$$\begin{aligned} p_{ij} &= P_r[X_1=j \mid X_2=i] \\ &= a_1^{\,j}(1-a_1)^{r-j}\cdot I_1^{r-i-j}\cdot I_3^{i}\cdot I_4^{j}\cdot \sum_{s=0}^{min(i,j)} {}_iC_s\cdot {}_{r-i}C_{j-s}\left[\frac{I_1 I_2}{I_3 I_4}\right]^s \end{aligned} \tag{2}$$

where

$$\begin{aligned} I_1 &= 1+\rho\sqrt{a_1a_2/\{(1-a_1)(1-a_2)\}}, \quad I_3 = 1-\rho\sqrt{a_1(1-a_2)/\{a_2(1-a_1)\}} \\ I_2 &= 1+\rho\sqrt{(1-a_1)(1-a_2)/(a_1a_2)}, \quad I_4 = 1-\rho\sqrt{a_2(1-a_1)/\{a_1(1-a_2)\}} \end{aligned} \tag{3}$$

By this direct expression of the conditional distribution using the some necessary parameters, we can easily calculate any probability and carry out a numerical simulation.

2) Marginal Distribution and Fundamental Statistics

Marginal distribution, mean, variance and skewness coefficient for X_i $(i=1,2)$ are given by the followings;

$$P_{X_i}(X_i) = {}_rC_{X_i}(1-a_i)^{r-X_i}(a_i)^{X_i},\ E(X_i) = r a_i$$

$$V(X_i) = r a_i(1-a_i),\ C_s = (1-2a_i)/\sqrt{r a_i(1-a_i)} \tag{4}$$

And the mutual correlation coefficient ρ is given by the non-negative value C,

$$\rho = (C - a_1 a_2)/\sqrt{a_1 a_2(1-a_1)(1-a_2)} \tag{5}$$

For such a correlation coefficient, the following constraint holds,

$$max\left[-\sqrt{a_1 a_2/\{(1-a_1)(1-a_2)\}},\ -\sqrt{(1-a_1)(1-a_2)/(a_1 a_2)}\right] < \rho <$$

$$min\left[\sqrt{a_1(1-a_2)/\{a_2(1-a_1)\}},\ \sqrt{a_2(1-a_1)/\{a_1(1-a_2)\}}\right] \tag{6}$$

3) Mean and Variance for Conditional Variate

Using eq.2, we can get the mean and variance of X_1 for a fixed value X_2. The result shows that the conditional mean is expressed by the next linear regression form to X_2 and that the conditional variance is also expressed by the linear relation to X_2.

$$\mu_1'(X_1|X_2) = \overline{X_1} + \rho\,(S_1/S_2)\left(X_2 - \overline{X_2}\right) \tag{7}$$

where, $\overline{X_i}$ and S_i denote the sample mean and the variance of data X_i $(i=1,2)$ respectively.

Because of this linear regression, the proposed distribution model is very useful for the estimation by regression curve.

4) Parameter Estimation by Moment Method

From eq.4, we obtain the shape parameter a_1 and a_2 by using the relation $a_i = 1 - V(X_i)/E(X_i)$ $(i=1, 2)$. From this, we get the upper boundary r.

$$r = 0.5 \times \left[E(X_1)/a_1 + E(X_2)/a_2\right] \tag{8}$$

Because the upper boundary r must be an integer, the obtained value should be converted into the common upper boundary $\hat{r}$ by rounding it to the nearest integer. Now $\hat{r}$ is an estimate of r. Then the shape parameter can be re-calculated by following relation:

$$\hat{a}_i = E(X_i)/\hat{r} \quad (i=1,2) \tag{9}$$

We can calculate the constants A, B_i, $(i=1,2)$ and C in the following steps. First, by adopting a sample correlation coefficient as a mutual correlation coefficient, we can estimate C by the following formula:

$$\hat{C} = \hat{a}_1 \cdot \hat{a}_2 + \hat{\rho}\sqrt{\{\hat{a}_1 \cdot \hat{a}_2 (1-\hat{a}_1)(1-\hat{a}_2)\}} \tag{10}$$

From eq.10, we know B_i by $\hat{B}_i = \hat{a}_i - \hat{C}$ and A is given by $\hat{A} = 1 - \hat{B}_1 - \hat{B}_2 - \hat{C}$.

Negative Binomial Distribution

By using the following replacement in the above result,

$$r = -k \ (k>0), \quad -(B_i + C) = a_i \quad (i=1,2) \tag{11}$$

We can obtain the theoretical relations for a negative binomial distribution. We describe the outline of results as follows:

1) Marginal Distribution and Fundamental relations

We can represent the marginal distribution by the following current form:

$$P_{X_i}(X_i) = {}_{X_i+k-1}C_{X_i} \cdot (p_i)^k \cdot (q_i)^{X_i} \quad (X_i = 0,1,2,\cdots) \tag{12}$$

where, $p_i = (1+a_i)^{-1}$ .and $q_i = a_i(1+a_i)^{-1}$ $(i=1,2)$

The fundamental statistics by the moment estimate are as follows:

mean $E(X_i) = k \cdot a_i$, variance $V(X_i) = k\,a_i(1+a_i)$ (13)

skewness coefficient $C_{si} = (1+2a_i)/\sqrt{k\,a_i(1+a_i)}$ (14)

correlation coefficient $\rho = (a_1 a_2 - C)/\sqrt{a_1 a_2 (1+a_1)(1+a_2)}$ (15)

2) Conditional distribution

The following formulation expresses the conditional distribution:

$$p_{i,j} = a_1^j (1+a_1)^{-k-j} I_1^{-k-i-j} I_3^i I_4^j$$

$$\times \sum_{s=0}^{\min(i,j)} \frac{\Gamma(k+i+j-s)}{\Gamma(k+i)} \frac{(-1)^s \cdot i!}{s!(i-s)!(j-s)!} \left[\frac{I_1 I_2}{I_3 I_4}\right]^s \tag{16}$$

where

$$I_1 = 1-\rho\sqrt{a_1 a_2/\{(1+a_1)(1+a_2)\}},\ I_3 = 1-\rho\sqrt{a_1(1+a_2)/\{a_2(1+a_1)\}}$$

$$I_2 = 1-\rho\sqrt{(1+a_1)(1+a_2)/(a_1 a_2)},\ I_4 = 1-\rho\sqrt{a_2(1+a_1)/\{a_1(1+a_2)\}}$$

The above correlation coefficient has the following constraint:

$$\rho < min\left[\sqrt{a_1(1+a_2)/\{a_2(1+a_1)\}},\ \sqrt{a_2(1+a_1)/\{a_1(1+a_2)\}}\right] \tag{17}$$

3) Parameter Estimation by Moment Method

By using the mean and variance, we can estimate the shape parameter a_i $(i=1,2)$ as $\hat{a}_i = V(X_i)/E(X_i) - 1$. $\hat{X}$ shows an estimate of X. We have k as the next approximation, $\hat{k} = 0.5\times\{E(X_1)/\hat{a}_1 + E(X_2)/\hat{a}_2\}$.

The shape parameter can be calculated by the following revised form, $\hat{a}_i = E(X_i)/\hat{k}$. By using the obtained results and the correlation parameter ρ estimated as a sample correlation coefficient, we can estimate C as:

$$\hat{C} = \hat{a}_1\hat{a}_2 - \rho\sqrt{\hat{a}_1\hat{a}_2(1+\hat{a}_1)(1+\hat{a}_2)}$$

Lastly B_i and A can be estimated by the following relations:

$$\hat{B}_i = -\hat{a}_i - \hat{C}\quad (i=1,2),\quad \hat{A} = \hat{B}_1 - \hat{B}_2 - \hat{C}$$

Moreover, you can easily get an anisotropic bi-variate geometric distribution by substituting $k=1$ in the above relations for the negative binomial distribution.

4) Parameter Estimation by Maximum Likelihood for Negative Binomial Distribution

Because the joint distribution is already known, we can get the parameter estimation by a maximum likelihood method in means of maximization of likelihood. However, the actual calculation is very difficult because of the constraints placed upon the correlation coefficient. Firstly we carry out the parameter estimation on marginal distributions and, secondly the correlation coefficient for joint distribution. We will illustrate the estimation only for the negative binomial distribution.

i) Estimation for The Marginal Distribution

In this case the unknown parameters are k_1, k_2, a_1 and a_2. We assume that the parameters k_1 and k_2 are different. In order to estimate these parameters, we firstly calculate the logarithmic likelihood LL on the basis of n joint data set $(X_{ij}: i=1,2,\cdots,n; j=1,2)$.

By the simultaneous equation $\partial LL/\partial k_j = 0$ and $\partial LL/\partial a_j = 0$, the maximum likelihood estimates k_j are given by the following formulations (6),

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{X_{ij}-1}\frac{1}{k_j+s} - ln\left(1+\frac{\overline{X_j}}{k_j}\right) \equiv D(k_j) = 0\quad (j=1,2) \tag{18}$$

The numerical value of the left side in the above equation means an increasing rate of LL by the increase of k_j per unit number of data. In practice, on the iterative adoption of k_j, when the absolute value of $D(k_j)$ exists within an allowable limit ε_a, the value k_j may be considered as an approximate solution. Then, the common parameter k is approximated by an average of k_1 and k_2. After we use these k_j $(j=1,2)$ parameters and eq.13, the shape parameter a_j is obtained by using the sample mean.

ii) Estimation of Correlation Parameter

Multiplying the marginal distribution by the conditional distribution leads to the joint distribution $f(i,j)=P_i \times p_{ij}$. This distribution has an unknown correlation parameter ρ. Then, we can try to estimate the necessary parameter by maximizing the logarithmic likelihood $LL=\sum_{i=1}^{n} log\, f(X_{i1},X_{i2})$ for the joint data set $(X_{i1},X_{i2}; i=1,2,\cdots,n)$ under consideration of the constraint.

NUMERICAL EXAMPLE BY THE ANISOTROPIC DISCRETE DISTRIBUTION

When we apply the model to the data in a season of low flow, we must pay attention to the runoff characters. We suggest that the adoption of sampling data may not be appropriate. The reason is that a few extreme data often lead to a wrong estimation of parameter. Parameter estimation should be done for each residual data removed 1st, 2nd,..., etc. using the upper sample range.

For checking the suitability of marginal distribution, we use the error term $D(k_j)$ and the Chi-square statistics. We have used the daily inflow data in Ayakita-Dam and Ayaminami-Dam basins, in Miyazaki Prefecture Japan. For our sample of the season of low flow, we have selected the time period of 120 days starting on 2nd November every year, for the years 1966-1986. We have adopted the unit of 3 days' duration and 1 $m^3/s/100km^2$ specific discharge. Time series of discharge are represented by means of a discrete number. These are shown by X_1 and X_2. Base-flow component should be removed in advance. Table 1 shows the outline of the procedure and the results.

When you exclude the upper range data, the average and standard deviation will usually decrease. Especially the decreasing rate of standard deviation is large. At first, the data fit a negative binomial distribution because the mean is smaller than the variance. Gradually the tendency will changes into (mean>variance). In other words the distribution tends toward a binomial distribution. In the table, the error term and Chi-square statistics show the goodness of fit. If we remove only few data, the Chi-square test may show a failure. If we remove the large amount of data, usually a goodness-of-fit test will result in success and the error function $D(k)$ will decrease. It is difficult to determine the limit of the allowable error term depended on the condition. By assuming that the allowable limit ε_a is 7×10^{-3}, the parameters k_1 and k_2 can be determined. The results, in a discrete integer form, are then, k_1 =2 and k_2 =5. We adopt the common parameter k=4 as a mean value.

By dividing each mean value by the common k, the estimates of shape parameter can be given by a_1=0.212 and a_2=0.229. The total sample number is 480 and that of partial samples is 460 for X_1 and 474 for X_2. Lastly, the correlation parameter is calculated by ρ=0.719 (maximum likelihood estimate) and 0.743 (moment estimate) for the joint sample of 456. For

Table 1 Numerical example of parameter estimation for an anisotropic bi-variate negative binomial distribution

cor. parameter	0.719 0.743	0.777 max. like. est. 0.829 moment est. mutual cor. coeff. ρ
parameter of marginal distr.	large $\alpha_1 = 0.212$ large $\alpha_2 = 0.229$	shape parameter small a_1, a_2
	large 4	small comon param. k
	small $\varepsilon_a = 7 \times 10^{-3}$	large est. error D(k)
	OK ⇦ ⇨	NO Chi-sq. test
	binomial distr. mean>variance small	⇦ ⇨ negative binomial distr. mean<variance large variance
	small	large mean
	partial	total adoped sample
marginal sample X_1 : 460, X_2 : 474 joint sample (X_1, X_2) : 459		480 number of adopted sample

the total sample, these results are 0.777 and 0.829.

We illustrate a comparison between an observed and a theoretical distribution. Fig.1 shows the comparison of the marginal distribution. There is a slight difference between them, but these data sufficiently fit this model. In the Chi-square test, a goodness of fit is acceptable at a confidence level about 1% for X_1 and 5% for X_2. Fig.2 shows the comparison of the joint distribution. In this case, as a whole, an approximate fitness can be explained adequately with this model.

CONCLUDING REMARKS

On reviewing past studies, we find that a discrete distribution with a mutual correlation has mainly be treated by an isotropic distribution. The above proposed anisotropic distribution model will be very useful for introducing flexibility of stochastic representation in the hydrological amounts. Especially we consider that the proposed distribution model will have a wide applicability of introducing on a mutual correlation in the two hydrological variables and also on a sequential correlation in considering the stochastic difference of successive time-series.

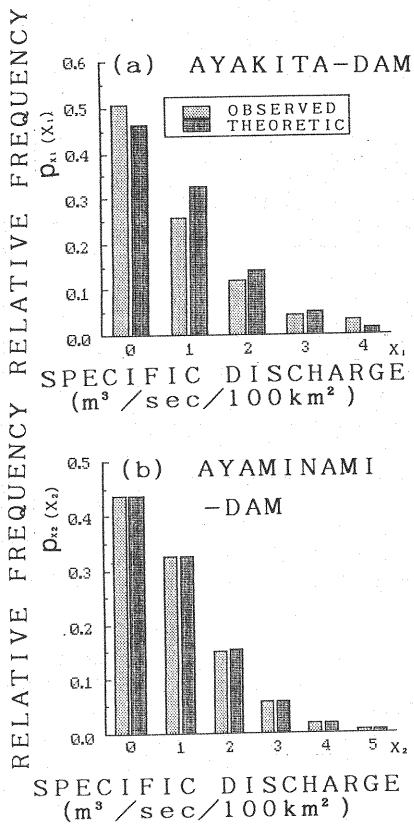

Fig.1 Comparison of thoretical and observed frequency of marginal distribution

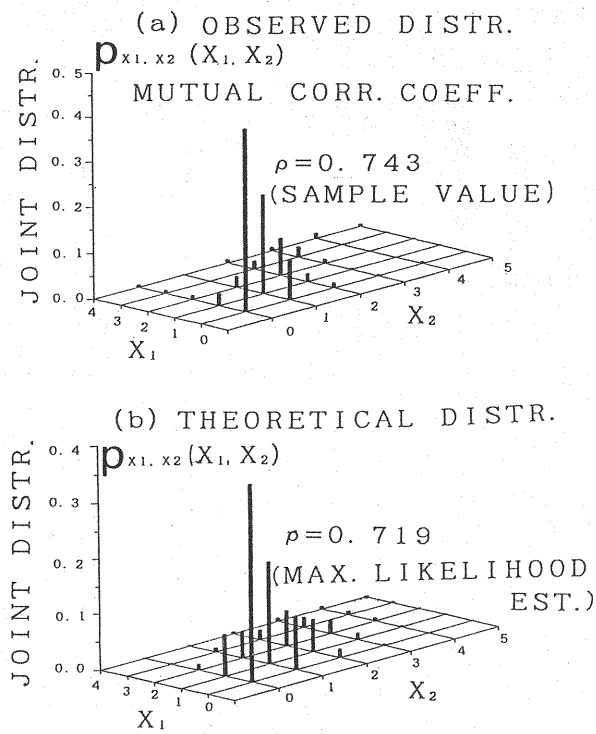

Fig.2 Comparison of theoretical and observed frequency of joint distribution

REFERENCES

1. Nagao, M. and Y. Ikeda : Stochastic process theory on reservoir function for supply with Markovian inputs, Proc. of 23th Japanese Conference on Hydraulics, pp.245-255, 1979 (in Japanese).
2. Nagao, M. and H. Konishi : A theoretical model of bi-variate binomial distribution with anisotropic shape parameters, Proc. of the annual conference of the JSCE, Chubu branch, pp.159-160, 1992 (in Japanese).
3. Nagao, M., H. Konishi, and M. Suzuki : Theory of anisotropic bi-variate discrete distribution --mainly bi-variate negative binomial distribution--. Proc. of the annual conference of the JSCE, II, pp.668-669, 1992 (in Japanese).
4. Nagao, M., H. Konishi and M. Suzuki : Anisotropic bi-variate discrete stochastic model for hydrologic series in dry season, Proc. of the 33th Japanese Conference on Hydraulics, pp.51-56, 1993(in Japanese).
5. Edwards, C.B. and J. Gurlands : A class of distributions applicable to accidents, Jour.

Amer. Statist. Ass., pp.503-517, 1961.
6. Johnson, N.L. and S. Kotz : Discrete Distribution, Houghton Mifflin Co., pp.131-132, 1966.

APPENDIX - NOTATION

The following symbols are used in this paper:

a_1, a_2, r	= shape parameter for variable X_1, X_2 and upper boundary;
$A, B_i, C \ (i = 1, 2)$	= constant value for p.g.f.;
${}_iC_j$	= binomial coefficient, that is, $i! / \{j! \cdot (i-j)!\}$;
$D(k_j)$	= error term for a maximum likelihood estimate of k_j;
$E(X_i), V(X_i), C_S$	= expectation, variance, skewness coefficient of X_i;
$f(X_i, Y_i)$	= joint distribution of X_i and Y_i
$G_{X_1,X_2}(Z_1, Z_2)$	= probability generating function (p.g.f.) of X_1 and X_2;
LL	= logarithmic likelihood function;
$max(A,B), min(A,B)$	= maximum, minimum value of A or B;
$p_{i,j}$	= conditional probability of $X_1 = j$ for a fixed value $X_2 = i$;
$P_{X_i}(X_i)$	= marginal distribution of X_i $(i = 1,2)$
$\overline{X}_i, S_i \ (i = 1,2)$	= sample mean, variance of data X_i $(i = 1,2)$;
ε_a	= allowable limit for maximum likelihood estimate;
$\mu_1'(X_1 \mid X_2)$	= conditional mean of X_1 for a fixed value X_2; and
ρ	= correlation coefficient between X_1 and X_2