An algorithm for generating positively correlated Beta-distributed random variables with known marginal distributions and a specified correlation

Steen Magnussen*

Canadian Forest Service, Natural Resources Canada, Pacific Forestry Centre, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada

Received 1 August 2002; received in revised form 1 June 2003

Abstract

A new algorithm for generating two positively correlated Beta-distributed random variables with known marginal distributions and a specified correlation is provided. The paired Beta-distributed random variables are generated from ratios of independent standard Gamma distributions. A positive correlation is achieved by introducing two shared standard Gamma distributed random variables. Parameters of the shared random variables are found by equating a one-term Taylor-series approximation of the expected covariance between the paired Beta random variables to the covariance commensurate with a specified correlation. The new algorithm is widely applicable. Upper limits of the correlation for given marginals are given. Tests of the algorithm revealed a small but persistent negative bias of 0.02 in achieved correlations. An application example is provided.

© 2003 Published by Elsevier B.V.

Keywords: Expected variance; Expected covariance; Standard Gamma distribution

1. Introduction

Generation of univariate random variables distributed according to some specified distribution is important for statistical analysis, inference and stochastic simulation (Robert and Casella, 1999; Devroye, 1986; Johnson et al., 1986). Simulation of time trends in a random variable is frequently done within a multivariate framework (Johnson, 1987)
where the different variables represent the state of the random variable at different points in time.

The flexible Beta distribution is widely used in the life sciences to describe the probability density distribution of proportions or relative frequencies of a random univariate variable (Massmann, 1981; Pitkanen, 1998; Collett, 1991). Generation of univariate Beta-distributed random variables is straightforward. Devroye (1986) and Johnson and Kotz (1970) provide standard algorithms based on ratios of standard Gamma-distributed or $\chi^2$-square distributed random variables, respectively. Generating pairs of correlated Beta-distributed random variables is less straightforward since there is no natural multivariate extension of the univariate Beta distribution (Johnson and Kotz, 1976). Although Plackett (1965) outlined the construction of bivariate distributions from known marginal distribution functions, and a known measure of ‘association’ the method, when tested, does not appear to apply to Beta distributions. Johnson (1987) found the Plackett method to work only for near normal and weakly correlated distributions. Loukas (1984) provided an algorithm for generating so-called bivariate Beta random variables. Yet, it is clear that the generated three parameter bivariate b-Beta distribution only has one marginal that is Beta-distributed, and that the distribution is restricted to a support domain where the sum of two random variables is less or equal to one. Also, the second variable is only Beta-distributed after a non-linear transformation. Michael and Schucany (2002) illustrated recently how the mixture approach could be used to generate bivariate Beta-distributed random variables with a specified correlation. Starting with a draw ($p_1$) from a Beta-distribution (prior) with known parameters $\alpha$ and $\beta$, a random integer number ($k$) of ‘successes’ is then drawn from a binomial distribution (likelihood) with parameters $p_1$ and $\nu$, finally a random draw ($p_2$) from a Beta-distribution with parameters $\alpha+k$ and $\beta+\nu-k$ will produce paired Beta-distributed random variables with a specified correlation of $\nu \times (\nu + \alpha + \beta)^{-1}$. The sequential nature of the algorithm, the relationship between the two sets of Beta-parameters, and the dependency of the correlation on $\alpha$ and $\beta$ clearly limits the application domain of this algorithm to situations where only the distribution of $p_1$ and the specified correlation is specified. The evolution from $p_1$ to $p_2$ is completely determined by $\alpha$, $\beta$, $k$, and $\nu$.

An algorithm for generating positively correlated random Beta variables without the above limitations does not seem to be currently available. This study propose an algorithm based on a first-order Taylor series expansion of the covariance of two Beta-distributed random variables when they share two standard Gamma-distributed variables. The algorithm is tested in a large number of settings, and an example of application in the context of evaluating the efficiency of various one-stage cluster sampling designs for estimating change in forest land cover type proportions is given.

2. Generating positively correlated random Beta variables

A random variable $X$ with a standard Beta probability density distribution having parameters $\alpha$ and $\beta$ ($Be(\alpha, \beta)$) can be generated from the following algorithm
Likewise, the complement to $X$ as a standard Gamma distribution having parameters $\Gamma(\alpha, 1)$ and the variance of a standard Gamma distribution with parameters $\Gamma(1, \beta)$ and $\Gamma(1, \beta')$ ensures that the marginal distributions of $X$ and $X'$ respectively. At time 1 $X_1$ is distributed as $\text{Beta}(\alpha_1, \beta_1)$ and at time 2 $X_2$ is distributed as $\text{Beta}(\alpha_2, \beta_2)$. Accordingly the expected value ($E$) of $X_1$ is $p_1 = \alpha_1 /(\alpha_1 + \beta_1)$ while that of $X_2$ is $p_2 = \alpha_2 /(\alpha_2 + \beta_2)$. Variances of $X_1$ and $X_2$ are $p_1(1 - p_1)/(1 + \alpha_1 + \beta_1)$ and $p_2(1 - p_2)/(1 + \alpha_2 + \beta_2)$, respectively. A popular technique for the generation of correlated random variables is to introduce a shared random variable (Devroye, 1986), a technique that extends to correlated binary variables (Park et al., 1996). The fact that the sum of two independent random standard Gamma variables, say, $\Gamma(\alpha_1, \beta_1)$ and $\Gamma(\alpha_2, \beta_2)$, is distributed as $\Gamma(\alpha_1 + \alpha_2, \beta_1 + \beta_2)$ makes the introduction of a shared Gamma random variable in the context of generating correlated Beta random variables straightforward. The shared random variables are created by a decomposition of the standard Gamma random variables in (1) into a sum of two independent random standard Gamma variables. Specifically

\[
X_1 = \frac{G(\alpha_1^* + \delta_1)}{G(\alpha_1^* + \delta_1) + G(\beta_1^* + \delta_2)},
\]

\[
X_2 = \frac{G(\alpha_2^* + \delta_1)}{G(\alpha_2^* + \delta_1) + G(\beta_2^* + \delta_2)},
\]

\[
1 - X_1 = \frac{G(\beta_1^* + \delta_2)}{G(\alpha_1^* + \delta_1) + G(\beta_1^* + \delta_2) + G(\delta_2)},
\]

\[
1 - X_2 = \frac{G(\beta_2^* + \delta_2)}{G(\alpha_2^* + \delta_1) + G(\beta_2^* + \delta_2) + G(\delta_2)}.
\]

Imposing the constraints $\alpha_1^* + \delta_1 = \alpha_1$, $\alpha_2^* + \delta_1 = \alpha_2$, $\beta_1^* + \delta_2 = \beta_1$, and $\beta_2^* + \delta_2 = \beta_2$ ensures that the marginal distributions of $X_1$ and $X_2$, and the covariance between $X_1$ and $1 - X_1$ and between $X_2$ and $1 - X_2$ remain unaffected by the introduction of the shared Gamma random variables $\delta_1$ and $\delta_2$. From (2) it is clear that $X_1$ and $X_2$ and their complements both share the random variables $G(\delta_1)$ and $G(\delta_2)$. This sharing creates a covariance between $X_1$ and $X_2$ and conversely between $1 - X_1$ and $1 - X_2$.

A method of moments solution for $\delta_1$ and $\delta_2$, in (2) can be obtained by requiring that the expected covariance between $X_1$ and $X_2$ and between $1 - X_1$ and $1 - X_2$ matches a target covariance (viz. correlation).
A first-order Taylor series approximation to the covariance between \(X_1\) and \(X_2\) and between their complements is

\[
\text{Cov}(X_1, X_2) = E\left( \frac{dX_1}{G(\delta_1)} \times \frac{dX_2}{G(\delta_1)} \right) \times \text{Var}(G(\delta_1))
\]

\[
+ E\left( \frac{dX_1}{G(\delta_2)} \times \frac{dX_2}{G(\delta_2)} \right) \times \text{Var}(G(\delta_2)),
\]

\[
\text{Cov}(1 - X_1, 1 - X_2) = E\left( \frac{d(1 - X_1)}{G(\delta_1)} \times \frac{d(1 - X_2)}{G(\delta_1)} \right) \times \text{Var}(G(\delta_1))
\]

\[
+ E\left( \frac{d(1 - X_1)}{G(\delta_2)} \times \frac{d(1 - X_2)}{G(\delta_2)} \right) \times \text{Var}(G(\delta_2)),
\]

where \(E\) denotes the expected value operator. Inserting given or estimated values of \(\alpha_1, \beta_1, \alpha_2, \beta_2\), and noting that \(E(G^2(x)) = \Gamma(2 + \alpha)/\Gamma(\alpha)\) the covariance expressions in (3) simplify to

\[
\text{Cov}(X_1, X_2) = \frac{\alpha_1 \times \alpha_2 \times \delta_2 + (1 + \beta_1) \times (1 + \beta_2) \times \delta_1}{(\alpha_1 + \beta_1) \times (\alpha_2 + \beta_2) \times (1 + \alpha_1 + \beta_1) \times (1 + \alpha_2 + \beta_2)},
\]

\[
\text{Cov}(1 - X_1, 1 - X_2) = \frac{\beta_1 \times \beta_2 \times \delta_1 + (1 + \alpha_1) \times (1 + \alpha_2) \times \delta_2}{(\alpha_1 + \beta_1) \times (\alpha_2 + \beta_2) \times (1 + \alpha_1 + \beta_1) \times (1 + \alpha_2 + \beta_2)}.
\]

Note the symmetry in the covariance expressions in (4) one is transformed to the other by replacing \(\alpha\) with \(\beta\) and \(\delta_1\) with \(\delta_2\) and vice versa. Equating \(\text{Cov}(X_1, X_2)\) to the known or estimated covariance \(\rho(X_1, X_2) \times \sqrt{\text{Var}(X_1) \times \text{Var}(X_2)}\) and \(\text{Cov}(1 - X_1, 1 - X_2)\) to \(\rho(X_1, X_2) \times \sqrt{\text{Var}(1 - X_1) \times \text{Var}(1 - X_2)}\) (since \(\rho(X_1, X_2) = \rho(1 - X_1, 1 - X_2)\)) and solving for \(\delta_1\) and \(\delta_2\) gives

\[
\delta_1 = \rho(X_1, X_2) \times (1 + \alpha_1 + \alpha_2) \times C,
\]

\[
\delta_2 = \rho(X_1, X_2) \times (1 + \beta_1 + \beta_2) \times C,
\]

\[
C = \frac{\sqrt{\alpha_1 \times \alpha_2 \times \beta_1 \times \beta_2 \times (1 + \alpha_1 + \beta_1) \times (1 + \alpha_2 + \beta_2)}}{(1 + \alpha_2) \times (1 + \beta_1) \times (1 + \beta_2) \times (1 + \alpha_1 + \beta_1 + \beta_2 + \alpha_2 (1 + \beta_1 + \beta_2))}.
\]

(5)

Solutions must further satisfy \(\delta_1 \leq \min(\alpha_i, \alpha_i)\) and \(\delta_2 \leq \min(\beta_i, \beta_i)\) to ensure \(\alpha_i - \delta_i \geq 0\) \(i = 1, 2\) and \(\beta_i - \delta_2 \geq 0\) \(i = 1, 2\). Consequently the maximum temporal correlation for a given pair of marginal Beta distributions is

\[
\max\{\rho(X_1, X_2)\} = \min\left\{ \frac{\alpha_1}{1 + \alpha_1 + \alpha_2}, \frac{\alpha_2}{1 + \alpha_1 + \alpha_2}, \frac{\beta_1}{1 + \beta_1 + \beta_2}, \frac{\beta_2}{1 + \beta_1 + \beta_2} \right\} C^{-1}.
\]

(6)
3. Testing the algorithm

The solution provided in (5) is based on a first-order Taylor series expansion; thus the achieved correlation between $X_1$ and $X_2$ may be biased. The performance of the algorithm was therefore tested for a series of Beta distributions and target values of $\rho(X_1,X_2)$ deemed realistic in the context of land use change studies. Beta distributions for $X_1$ were generated for $x_1 = 1, 3, 5, \ldots, 31$, and $\beta_1 = x_1, x_1 + 10, x_1 + 20, \ldots, x_1 + 100$ for each $x_1$. The corresponding distributions for $X_2$ were generated by replacing $x_1$ with $\gamma \times x_1$ where $\gamma = 0.8, 1.2$ and $\beta_1$ with $\eta \times \beta_1$ where $\eta = 0.9, 1.1$ for a total of $704 = 16 \times 11 \times 2 \times 2$ paired distributions. Target correlation for each pair of Beta distributions was uniform in $[0.2, 0.9]$.

Results of the testing are shown in Figs. 1–3. Fig. 1 shows the achieved (empirical) correlation coefficient $\rho_{\text{emp}}$ plotted against the target correlation $\rho_{\text{target}}$. The agreement is generally good with an adjusted $R^2$ of 0.988 and a least squares linear relationship of $\hat{\rho}_{\text{emp}} = -0.0041 + 0.982 \times \rho_{\text{target}}$ indicating a systematic downward bias in the achieved correlations. A histogram of residuals $\rho_{\text{emp}} - \rho_{\text{target}}$ is in Fig. 2, confirming the preponderance of a slight negative bias. Fig. 3 provides a contour plot of the (negative) relative bias for given values of the parameters of the Beta distribution for $X_1$. It is apparent that the negative bias is small ($\approx 5\%$) for a large range of parameter values.

Fig. 1. Scatterplot of achieved ($\rho_{\text{emp}}$) versus specified ($\rho_{\text{target}}$) correlation between paired random variables with known marginal Beta-distributions.
for $X_1$, while a serious bias ($> 10\%$) is restricted to near exponential or exponential-type Beta distributions (inverse $J$-shaped) for $X_1$.

4. An application example

The need for generating correlated beta distributed random variables arise, for example, in the context of gauging the efficiency of various one-stage cluster sampling designs for estimating temporal change in a categorical variable (Cochran, 1977). The following example illustrates an application of the above algorithm during an assessment
of various design options for Canada’s new national forest inventory (Magnussen, 2001; Magnussen et al., 1998). Analysis of historic regional forest cover type maps indicated that the proportion of hardwood forests at time $t_1 (X_1)$ in 2 km × 2 km primary sampling units could be described as $Be(5,30)$ with $E(X_1) = 0.143$, and $Var(X_1) = 0.0340$. The corresponding estimates for time $t_2$ 10 years later ($X_2$) were $Be(4,31)$, $E(X_2) = 0.114$, and $Var(X_1) = 0.0281$. The temporal correlation of proportions in the 2 km × 2 km plots was 0.82. From (5) one obtains $\delta_1 = 3.608$ and $\delta_2 = 22.377$. According to (6) the maximum possible temporal correlation is estimated at 0.91. With these solutions 12000 paired random realizations of $X_1$ and $X_2$ were generated as per (2) and assigned at random to a primary sampling unit in the region. The efficiency of various sampling designs to estimate decadal change in the relative hardwood cover of the region was then assessed by simulated sampling of primary sampling units with paired values of $X_1$ and $X_2$ having the prescribed marginal

![Graphical representation](image)

**Fig. 4.** Relative frequency distribution of 12000 simulated values of correlated Beta distributed random variables $X_1$ and $X_2$ (gray bars). The known specified Beta distributions ($Be(5,30)$ and $Be(4,31)$) are superimposed (black lines).
Fig. 5. Scatterplot of a random subset of size 300 taken from the 12 000 random realizations of $X_1$ and $X_2$ summarized in Fig. 1.

distribution and the specified temporal correlation. Histograms of the relative frequencies of $X_1$ and $X_2$ are displayed in Fig. 1 together with the expected Beta probability density distributions. The maximum absolute differences between the simulated and expected marginal distributions was 0.01. The empirical correlation coefficient of $X_1$ and $X_2$ was 0.79, slightly below the target of 0.82. Fig. 2 shows a scatterplot of a random subset of 300 paired random realizations of $X_1$ and $X_2$ (see Figs. 4 and 5).

5. Discussion and conclusions

The proposed algorithm extends the method of Michael and Schucany (2002) to applications where both marginal distributions and the correlation are specified. The fact that the specified correlation between two random Beta variables with known distributions has an upper limit is logical and also implicit in the Michael and Schucany algorithm. The shared standard Gamma variable can be viewed as the invariable part of a countable measure that cannot exceed the whole part of any measure. There exist, thus, an upper limit to the temporal correlation for a given pair of Beta-distributed random variables that is independent of the generating algorithm. Part of the negative bias in the achieved correlation is due to attenuation by a lack of fit in the specified marginal distributions (Crowder, 1985; Mosimann, 1962; Yaglom, 1986). Attenuation increased, as expected, with the skewness coefficient of the two marginal Beta
distributions. The remainder of the bias is due to non-vanishing higher terms of the Taylor-series expansion of the expected covariance.

As illustrated, the new algorithm allows the simultaneous generation of random variables with known marginal Beta distributions and a specified correlation. Statistical inference on linear combinations of these random variables follows standard procedures (Casella and Berger, 2002). A major application domain for the proposed algorithm is expected for statistical inference on temporal changes in land use and land cover (Corona et al., 2002, Anderson, 2002, Schreuder et al., 1999) and would include estimation of model-based bootstrap confidence intervals (Efron and Tibshirani, 1993) for the difference of correlated proportions (Lloyd, 1999; Agresti and Caffo, 2000).

References


