Several generation methods of multinomial distributed random number

Tian Lei[1, a] , LinxiHe[1,b],Zhigang Zhang[1,c]

[1] School of Mathematics and Physics, USTB, Beijing 100083, China

[a] carolineleitian@163.com, [b] helinxi@163.com, [c]zzgcyf@263.net

Keywords: multinomial distribution, pseudo random number, Gibbs sampling, simulation

Abstract. The paper introduces three different ways to generate pseudo random number of multinomial distribution and also describes the implementation process in detail with the model of small ball and box. We mainly consider the generation using Gibbs sampling in MCMC method. In the paper, we intend to find the high-quality generation of pseudo random number of multinomial distribution.

Introduction

To some extent, multinomial distribution can be considered as a generalization of the binomial distribution. Bernoulli trials of binomial distribution only have two opposite possible outcomes A_1、$A_2$and the result of each trail is independent.

In Bernoulli trials of binomial distribution, let random variables x stand for the results of trials, p stands for the probability of the event occurs. So, in $n-$timestrial, the probability of k times occurrenceis $P(x=k)=C_n^k p^k(1-p)^{n-k}$,so that $X \sim B(n,p)$.

In multinomial distribution, there are several results$A_1, A_2, \cdots, A_m, A_{m+1}$and their probabilities are $p_1, p_2, \cdots, p_m, p_{m+1}$which satisfy $\sum_{i=1}^{m+1} p_i = 1$. Therefore, the probability thatA_1 occurs x_1 times, A_2 occurs x_2 times, … , A_m occurs x_m times, A_{m+1} occurs x_{m+1} times,is $\frac{n!}{x_1!...x_{m+1}!} p_1^{x_1} p_2^{x_2} ... p_m^{x_m} p_{m+1}^{x_{m+1}}$ and $\sum_{i=1}^{m+1} x_i = n$.

Therefore, like binomial distribution, we define the joint distribution principle and other equations in multinomial distribution, thus, we get the following formulas.

$$P\{X_1 = x_1, \cdots X_m = x_m\} = C_n^{x_1} C_n^{x_2} ... C_n^{x_m} C_n^{n-(x_1+...+x_m)} p_1^{x_1} p_2^{x_2} ... p_m^{x_m} (1-(p_1+...+p_m))^{n-(x_1+...+x_m)}$$

$$= \frac{n!}{x_1! x_2! ... x_m! (n-(x_1+...+x_m))!} p_1^{x_1} p_2^{x_2} ... p_m^{x_m} (1-(p_1+...+p_m))^{n-(x_1+...+x_m)} \quad (1)$$

$$\sum_{i=1}^{m+1} p_i = 1, p_{m+1} = 1 - \sum_{i=1}^{m} p_i \text{ and } \sum_{i=1}^{m+1} x_i = n, x_{m+1} = n - \sum_{i=1}^{m} x_i$$

Hence, we define $X = (X_1, ..., X_m) \sim M(n, m, [p_1, p_2, ... p_m])$ as multinomial distribution.

Generations of pseudo random number of multinomial distribution

When studying the multinomial distribution, we consider the problem of throwing n small balls into m boxes. The probabilities of the small ball into the $i-th$ box separately are p_1、$p_2 \cdots p_m$. Hence, the probability satisfies the approximate formulas.

$$P(X_1 = x_1, X_2 = x_2, ..., X_{m-1} = x_{m-1})$$

$$= \frac{n!}{x_1! x_2! ... (n-(x_1+...+x_{m-1}))!} p_1^{x_1} p_2^{x_2} ... (1-(p_1+...+p_{m-1}))^{n-(x_1+...+x_{m-1})} \quad (2)$$

However, when the numbers of the small balls and boxes are large, there is high probability of boundary violation in computer's calculation because of the large factorial term. Hence, we apply the logarithmic transformation in the above formulas.

© 2015. The authors - Published by Atlantis Press

$$\ln P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m) = \ln(\frac{n!}{x_1!x_2!...x_m!} p_1^{x_1} p_2^{x_2} ... p_m^{x_m})$$
$$= (\ln n + \ln(n-1) + ... + \ln 1) - \sum_{i=1}^{m}[\ln(x_i) + \ln(x_i - 1) + ... + \ln 1] + \sum_{i=1}^{m} x_i \ln p_i \quad (3)$$

In the simulation, we choose the experiment putting 20 small balls into 4 boxes 100 times. In the experiment, P_1,the probability that small balls fall into the first box each time is 0.1. Analogously, we can define that $P_2 = 0.2, P_3 = 0.3, P_4 = 0.4$.

Separation into multidimensional 0-1 vectors

Divide $X = (X_1, X_2, ..., X_m)$ into X_i, which represents the number in the $i-th$ box. Hence, we separate X into n 0-1 random vectors with m dimens. We put n small balls into boxes one by one. Define 0-1 random vectors $Y_j = (y_{j1}, y_{j2}, ..., y_{j,m})$, Y_j as the result of the $j-th$ ball. And, the value of $y_{j1}, ...y_{j,m}$ is only 0 or 1 and y_{j1}+...+$y_{j,m} \leq 1$. For example, if y_{j1}=1, it shows that the $j-th$ ball falls into the first box. According to the definition, we can get $X_i = \sum_{j=1}^{n} y_{ij}$. Because each ball falls into boxes independently, Y_j is independent from each other and $P(Y_j) = (p_1, p_2, ..., p_m)$.

Using Inverse Transformation Method, we can determine which box the small ball falls intoaccording to the uniform random numbers. Repeating the trail n times we can simulate the multinomial distribution.

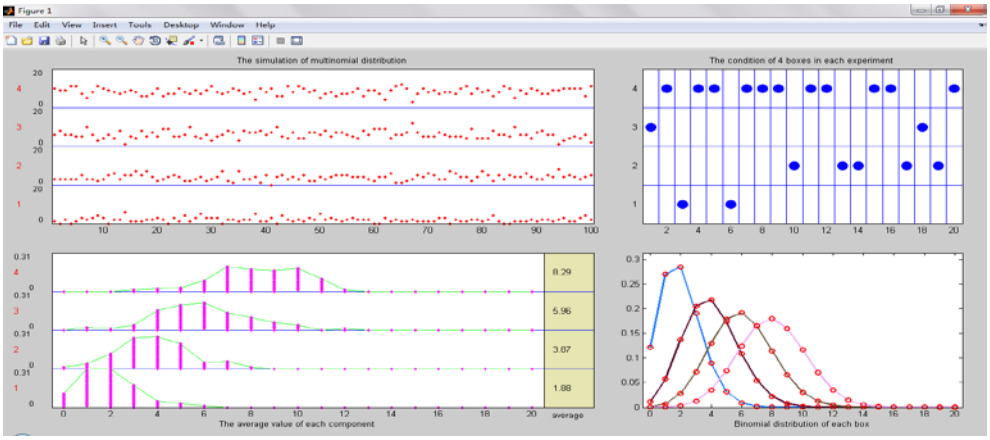

Fig. 1: The simulation according to separation into multidimensional 0-1 vectors.

InFig. 1, the abscissaof the top left diagramrepresents 100 experiments and the value of the ordinate separately stands for the number of small balls in one of these 4 boxes.The diagram in the top leftdynamically displays the number of small balls in each box in each experiment. The diagram in the lower left shows the distribution law of small balls in each box, the frequency and the average value of number of different balls appearing in each box in 100 experiments. The diagram in the lower right displays the situation of binomial distribution in each box with given distribution probability.

Multiplication formula of the joint distribution

The multiplication formula of the joint distribution:

$$P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)$$
$$= P(X_1 = x_1) \cdot P(X_2 = x_2 \mid X_1 = x_1) \cdot P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2) \cdot ... \cdot P(X_m = x_m \mid X_1 = x_1, ..., X_{m-1} = x_{m-1}) \quad (4)$$

In the Eq.4. $\sum_{i=1}^{m} x_i = n$

The conditional probability formula:

$$P(X_2 = x_2 \mid X_1 = x_1) = \frac{P(X_2 = x_2, X_1 = x_1)}{P(X_1 = x_1)} = \frac{\frac{n!}{x_1!x_2!(n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2}}{\frac{n!}{x_1!(n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1}} \quad (5)$$

$$= \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} (\frac{p_2}{1-p_1})^{x_2} (\frac{1-p_1-p_2}{1-p_1})^{n-x_1-x_2}$$

i.e. $X_2 \mid X_1 = x_1 \sim B(n-x_1, \frac{p_2}{1-p_1})$

According to the formulas (5), we can derivate the formulas below.

$$P(X_m = x_m \mid X_1 = x_1, ..., X_{m-1} = x_{m-1}) = \frac{P(X_1 = x_1, ..., X_{m-1} = x_{m-1}, X_m = x_m)}{P(X_1 = x_1, ..., X_{m-1} = x_{m-1})}$$

$$= \frac{\frac{n!}{x_1!x_2!...x_{m-1}!(n-(x_1+...+x_{m-1}))!} p_1^{x_1} p_2^{x_2} ... p_{m-1}^{x_{m-1}} (1-(p_1+...+p_{m-1}))^{n-(x_1+...+x_{m-1})}}{\frac{n!}{x_1!x_2!...x_{m-2}!(n-(x_1+...+x_{m-2}))!} p_1^{x_1} p_2^{x_2} ... p_{m-2}^{x_{m-2}} (1-(p_1+...+p_{m-2}))^{n-(x_1+...+x_{m-2})}} \quad (6)$$

$$= \frac{(n-(x_1+...+x_{m-2}))!}{x_{m-1}!(n-(x_1+...+x_{m-1}))!} (\frac{p_{m-1}}{1-(p_1+...+p_{m-2})})^{x_{m-1}} (\frac{1-(p_1+...+p_{m-1})}{1-(p_1+...+p_{m-2})})^{n-(x_1+...+x_{m-1})}$$

i.e. $X_m \mid X_1 = x_1, ..., X_{m-1} = x_{m-1} \sim B(n-(x_1+...+x_{m-2}), \frac{p_{m-1}}{1-(p_1+...+p_{m-2})})$

Firstly, we consider the situation of the first box. When the number of small balls in the first box is determined, we consider the situation of the second box. Like this, we can determine the number of small balls in the second and the third box separately. Finally, we put all remaining balls into the last box. According to the definitions above,random vector satisfies $X_1 \sim B(n, p_1)$ and the conditional probabilities of random vectors X_2, X_3 satisfy binomial distribution. Using the sampling in binomial distribution, we can get values x_1, x_2, x_3 and $20-(x_1+x_2+x_3)$ is the number of small balls in the fourth box.

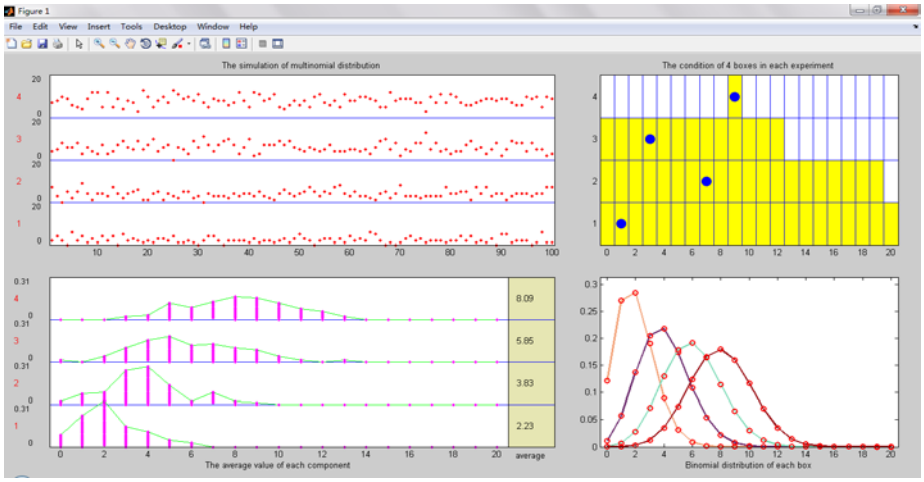

Fig. 2: The simulation according to multiplication formula of the joint distribution.

When we compareFig. 2 with Fig.1, the difference is that the grid meshes filled with yellow shows the number of small balls in each binomial distribution. In the result of the simulation, we can see each dimension of multinomial distribution still satisfies the binomial distribution.

Gibbs sampling

Using Gibbs sampling, we can get an appropriate limiting probability distribution of Markov chain. The initial condition x_1, x_2, x_3, x_4 is set to any random vector including four non-negative integers with the sum of them is 20. And x_i represents the number of the balls in the $i-th$ box. Then, we change conditions in the following way, choosing two numbers from $1,2,\cdots,i$ randomly to determine the next condition. For example, if we choose 1， 3, let $s = x_1 + x_3$ and keep the value of x_2, x_4 constant to simulate the new value of x_1, x_3. We control one of the variables x_1, x_3 to determine the other. We can presume that variable $x_3 \sim B(s, \frac{p_s}{p_1 + p_s})$ and $x_3 = 1,...,s-1$. According to inverse discrete transformation method, we can get the value of variable x_3 as v so that x_1 is $s-v$. Hence, we obtain the value of x_1, x_2, x_3, x_4 in the next condition.

If we allow the box is empty, the number of balls in each box can be 0. Firstly, we use the following method to obtain the initial condition x_1、x_2、x_3、x_4:

1) Generate a integer random number uniformly distributed in $[0,20]$ as x_1.
2) Generate a integer random number uniformly distributed in $[0,20-x_1]$ as x_2.
3) Generate a integer random number uniformly distributed in $[0,20-x_1-x_2]$ as x_3.
4) Let $x_4 = 20 -$ （$x_1+x_2+x_3$）.

After we get the initial condition, let x_1、x_2、x_3、x_4 as the first column of array X. In the simulation using Gibbs sampling introduced in section 2.3, we set number realization as 100 and track number as 20 for more stable results.

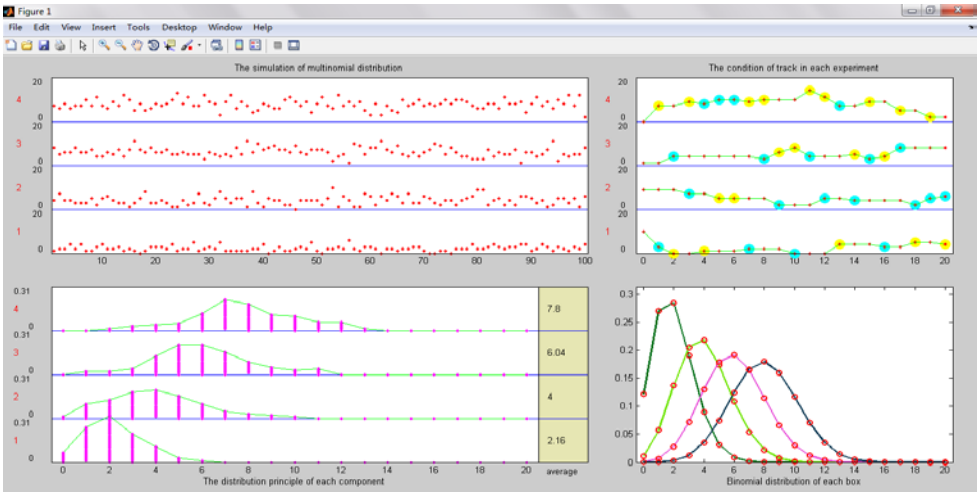

Fig. 3: The simulation according to Gibbs sampling.

Comparing with otherfigures above, the diagram in the top right ofFig. 3is quite different from others. These yellow small circles and green small circle stand for the two boxes, which are chose to change their value in this track. The abscissa represents track number 20, in other words, we get a stable result after 20 exchanges.

Acknowledgements

This work was financially supported by the Innovation Program of University of Science and Technology Beijing(15210067)

Reference:

[1]Chen LanXiang, F 一Minimax Estimators of the Parameters of the multinomial distribution, Journal of Mathematical Research and Exposition,Vol.9,No.4, p.489-494(1989)

[2]ShiFeng Xiong, GuoYing Li: Test of the Maximum Probability of the multinomial distribution, SCIENCE IN CHINA Ser. A Mathematics2005 No. 07，p.816-830,(2005)
[3] Melnikova, Y :History Matching Through a Smooth Formulation of Multiple-Point Statistics，MATHEMATICAL GEOSCIENCES，Vol 47, No4,p.397-416(2015)
[4] Baker, S.,Cousins, R.D:Clarification of the use of chi-square and likelihood functions in fits to histograms,Nuclear Instruments & Methods in Physics Research, Section A，Vol.221,No.2，p. 437-442 ,(1984)
[5] Hadfield, JD: MCMC Methods for Multi-Response Generalized Linear Mixed Models,JOURNAL OF STATISTICAL SOFTWARE,Vol.33,No.2,p.1-22,(2010)
[6] HU Yue:Multinomial distribution and multi-Poisson distribution, Journal of Zhejiang University of Science and Technology,Vo l.17 No .3,(2005)