Sindh Univ. Res. Jour. (Sci.Ser.), Vol: 34-B (2) 1 -10 (2002)

SIMULATION OF DISCRETE UNIVARIATE AND MULTIVARIATE PROBABILITY DISTRIBUTIONS

Muhammad Aslam

Department of Statistics, University of Karachi, Pakistan

Abstract

Simulation describes a wealth of varied and useful techniques, all connected with the mimicking of the rules of a model of some kind. (Shannon,1975). In this article an exploratory technique for the simulation of discrete univariate and multivariate distribution is proposed. The technique used only the random number table (Scott & Lindley, 1984). The uniform random numbers are used for simulation of randon experiment which would be modelled by the discrete probability distribution with certain parameters.

In this article it is shown that the technique may be used to simlate the following distributions: namely, Bernoulli, Binomial, Negative binomial, Geometric & Hypergeometric (Hogg and Craig, 1978). However, in the examples, only binomial, geometric and trinomial are discussed. The adequacy of the model was tested for the empirical probability distribution by using the Chi-square test statistics (Hogg & Craig, 269-1978). The exploratory method is also used to judge the quality of simulated values of random variables graphically (Hoaglin, Mosteller & Tukey, 1985).

Keywords: Simulation, Probability Distributions, Linear congruential generator, Chi-square test, Exploratory plots

Introduction

Simulation has been defined as the process of designing a model of a real system and conducting experiments in this model for the purpose either of understanding the behaviour of the system or evaluating various strategies within the limits imposed by a certain set of criteria for the operation of system.

The classical method for the simulation of discrete random variables known as table-look-up method (Morgan p-91, 1987). This method simulates the values of random variables X taking values 0,1,2 ..... etc. With probabilities; $p_i = P(X = i)$ for $i > 0$. First of all construct the cumulative probability distribution of X;
and select a uniform random number from the random number table (Scott & Lindley, 1984). The value of X corresponding to the upper limit of probability interval in which the selected random number lands become the simulated value of random variable X. The other method for simulation of discrete distribution is also discussed by the following researchers: Hormann (1993a) used transformed rejection method for Poisson, binomial distributions, he gives short and fast algorithms, Stadlober (1989) used the ratio of uniform method and Kachitrichanukul (1985) discussed a general method based on transformed density technique which has been applicable to discrete as well as continues distributions.

In the present method the basic tool is again the uniform random variable taken from radom number table (Scott & Lindley, 1984) or by using linear

CamScanner

congruential pseudo-random number generators (LCG) (Morgan, p-61, 1987; Hormann p-96-106, 1994). We want to simulate any random experiment which would be modelled by the probability distribution of discrete random varibale "X", with certain parameters. In the method discussed in this article we have no need to construct the cumulative probability distribution in contrast to table-up method but we have to partition the random numbers into number of disjoint subset, and the proportion of total number in each subset should match the probability of occurrence of the particular outcome of random experiment.

The technique of simulation discussed in this article also have been applied to bivariate (Trinomial distribution).

2. Method

The random number table (Scott & Lindley, 1984) is the basic tool of simulation of diserete random variable. The random number table containing the number generated by using the ten digits i.e. 0.1,3,4,5,6,7,8,9; which are having equal chances of selection. Generally we have to partition the set of ten digits into number of disjoint subsets such that each subset containing the particular random number digits which are proportional to the chances of occurrence of a particular outcome of random experiment. Then we copy down "K" random number/digits to simulate "K" values of any discrete random variable.

The exploratory method is also used to judge the quality of simulated values of random variables graphically (Hoaglin, Mosteller & Tukey, 1985). In this method the observed frequency "n_k" from any discrete probability distribution P (X_i) transformed into empirical expected frquency by using the following formulae:

$$\dot{n}_k = \begin{cases} n_k - 0.67 - (0.8 * n_k / K): \text{ for } n_k > 2 \\ 1 / exp\,(1); \quad ; \text{for } n_k = 1 \text{ ; where} \\ \text{"K"is total of frequency.} \\ Undefined; \quad ; \text{for } n_k = 0 \end{cases}$$

The theoretical expected frequency, that is; $n_k = KP(X_i)$ manipulated in the from of straight line. The plot is constructed by plotting the random variable values "X" versus count met meter that is $\overset{x}{\varphi}(n_k) = \log(n_k / K^n C_k) = n^* \text{ In } (I - p) + \text{In } (p/I\text{-}p))\, X$, for binomial distribution and versus count metameter that is $\varphi(\dot{n}_k) = \log(\dot{n}_k/K) = \text{Inp+In(I-p)}\, X$, for geometric distribution. {Hoaglin,Mosteller & Tukey, 1985).

3. Application

Now we discuss the techniques for the simulation of various discrete random variables.

Example 1 : Tossing a fair coin. (Binomial Distribution)

To simulate the values of Binomial random variable with probability of success "P", we have to divide the ten digits into two disjoint subset such that one subset containing the particular digits with the proportion of digits is "P" and other subset containing the digits with proportion of "(I - p)" corresponding to occurrence "success" & "failure" respectively.

To simulate the random experiment of tossing five fair coins 200 times and recording of the values of random variable X = number or head per trial. The outcome is "head" if the random digit selected from the random number table belong to the

set {0,2,4,6,8,} otherwise the outcome is tail i.e. random digit belongs to set {1,3,5,7,9}. We have to copy down 200 five-digital random numbers from the random number table and each digit of the five-digital random number shows the outcome of a coin, according to the above classification of digits assign the values of random variable X for each five-digital random number. E.g. if the random number is 49487, the value of random variable X equals 3 because the sequence of digits containing 3 digits belongs to the set {0,2,4,6,8}, corresponding to the occurrence of " head" as success.

Consider the following theoretical probability distribution of "X" empirical & expected frequency distributions given in table 1.

The value of chi-square statistics (Hogg & Craig p-269, 1978), comes out as 2.032 which is not very highly significant at P-value= 0.8447 with 5 degrees of freedom. (i.e. $2.032 \geq \chi^2_{(0.01,5)} = 15.086$). So we conclude that the empirical distribution adequately follows the binomial distribution with parameter n=5, p=0.5. The exploratory plot to check whether the simulated data is consistent with a binomial distribution i.e. b (n=5, p=0.5) is given in figure 1. The plot is constructed by plotting the binomial random variable values "X" versus count metameter that is $\varphi^*(n_k) = \log(n_k/Kn\ Ck) = n^* In (p/(I-p))$, (Hoaglin, Mosteller & Tukey, 1985). The plot indicates quite satisfactory agreement with the binomial model because the scatter of the plot is linear. An eye-fitted line yields a slope of roughly 0.02, so the estimate of p graphically found is 0.50499, this estimates of "p" slightly differs from the actual value of p = 0.5.

Example 2: Geometric distribution

Now we consider the simulation of random experiment in which the number of attempts required to qualify an aptitude test is a random variable with known probability of success 0.7. The random variable is defined as X=number of attempts (trials) to get through the test. We have to divide the set of ten random digits into two disjoint subsets such that, the outcome is "success" if the random digits selected from the random number table belongs to set {0,1,2,3,4,5,6} other wise outcome is "failure" if random digits belong to set {7,8,9}. To simulate the one value of random variable "X", we should continue to copy down the sequence of one digital random until we get exactly one digit belonging to the set {0,1,2,3,4,5,6} which corresponds to the "success" and count the number of random digit in the sequence as the value of random variable "X". For example consider the following sequence of random digits, 97973, the value of X = 5. Consider the following theoretical probability distribution, empirical & expected freqency distributions of "X" given in table 2, when 290 values of "X" are simulated by this method.

The value of chi-square statistic (Hogg & Craig p-269,1978) comes out as 2.64879, which is not significant at P-value=0.449 with 3, degrees of freedom (i.e. $2.64879 \not\geq \chi^2_{(0.01,3)} = 11.435$). So we conclude that the empirical distribution adequatelty follows the geometric probability distribution with parameter p = 0.7. The exploratory plot to check whether the simulated data is consistent with a geometric distribution with parameter P=0.7 is given in Figure 2. The plot is constructed by plotting the geometric random vaiable values "X" versus count metameter that is $\varphi^*(n_k) = \log(n_k/K) = Inp + In\ (p/(I-p))$, (Hoaglin, Mosteller & Tukey, 1985). The plot indicates quite

satisfactory agreement with the geometric model because the scatter of the plot is linear. An eye-fitted line yield a slope of roughly (-1.275), so the estimate of "p" graphically found is 0.720569. This estimates of "p" slightly differs from the actual value of p = 0.7.

Example: 3.
Multinomial Distribution

In industry the electronic equipment produced would be classified into three categories, namely "Satisfactory", Undersized and "Oversized" with probabilities $p_1 = 0.7$, $p_2 = 0.1$ & $p_3 = (1-p_1-p_2) = 0.2$. A random sample of size five units is selected from the production line. For simulation of 200 pairs of two random vaiables X_1 = Number of "satisfactory units" and X_2 = Number of " under sized'' units per sample, we have to partition the two-digital random numbers into two disjoint sets such that the unit would be classified as Satisfoctory" if the selected two-digital random number belongs to the set {70-79}. To simulate a pair of values of random variables (X_1,X_2) we should copy down the sequence of five (because the sample size of selected units is five) two-digital random and assign the values to the sequence of random numbers jointly to the random variables. Aaccording to the above calssification of random numbers. For example if the sequence of generated random numbers is 12,27,78,57,13 the corresponding values of random variables are $X_1 = 4$ & $X_2 = 1$.

Theoretical joint probability model of X_1 & X_2 is known as trinomial distribution with parameters n = 5, $p_1 = 0.7$ & $p_2 = 0.1$ given below (Hogg & Craig p-95, 1978):

$$P(x_1,x_2) = \frac{5!}{x_1!\, x_2!\, (5-x_1-x_2)!}(0.7)^{x1}\,(0.1)^{x2}\,(0.2)^{5-x1-x2}\ ;\ (x_1,x_2) = 0.1,\ldots 5\ ;\ \text{Provided that } (x_1+x_2) \le 5$$

Table 3 shows empirical & theoreical bivariate frequency distribution of (x_1,x_2).

The basic assumption to apply the chi-square statistics for the goodness of fit is that each expected frequency should be greater than five, in table 3 the highlighted expected cell frequencies are less than five, so we collapse the table 3 into table 4.

The calcualted value of chi-square statistics (Hogg & Craig p-269,1978) from Table 4 is 5.5.392 which is not significant at P-value = 0.69869 with 8 degrees of freedom (i.e. $5.5392 \ge \chi^2_{(0.01,8)} = 20.09$). Hence we conclude that the empirical joint probability distribution of (X_1, X_2) simulated in table-3 adequately (significantly) follow the trinomial distribution with parameters; $p_1 = 0.7$, $p_2 = 0.1$ & n =5. The discrepaney between the empirical and expected frequency obtained is shown graphically by the Three dimensional Multiple line chart in Figure 3, by using statistical software STATGRAPHIC (1988).

We know that the marginal probability distributions of X_1 and X_2 from the trinomial distribution follow the binomial with parameters n=5, $p_1 = 0.7$ and n=5, $P_2=0.1$ respectively. We would apply the chi-square statistics to test the adequacy of marginal distribution of X_1 and X_2 simulated at the right & bottom margin of Table 3, the value of chi-square statistics for the marginal distribution of X_1 comes out as 3.716755, which is not significant at P-value = 0.44569 with 4 degrees of freedom (i.e. $3.716755 \not\ge \chi^2_{(0.01,4)} = 13.277$). So we conclude that the empirical distribution of X_1 adequately follows the binomial distribution with parameter n=5, $p_1=0.7$.

The value of chi-square statistics for the marginal distribution of X_2 comes out as 0.07537, which is not significant at P-value = 0.96302 with 2 degrees of freedom ((i.e. 0.07537 $\geq \chi^2_{(0.01,2)} = 9.2103$). so we conclude that the empirical distribution of X2 adequately follows the binomial distribution with parameter, n=5, p_2=0.1.

Table 1: Observed and expected Frequency distributions and metameter of coin tossing experiment.

X_i	0	1	2	3	4	5
Observed Frequency n_k	06	33	68	55	30	08
Theoretical [rpnano;otu P (X_i)	1/32	5/32	10/32	1/32	5/32	1/32
Expected freq. $e_i=K*P(X_i)$	6.25	31.25	62.5	62.5	31.25	6.25
$\dot{n}_k$	5.306	32.198	67.058	54.110	29.210	7.298
Metameter $\dot{\varphi}\ (n_k)$	-3.629	-3.435	-3.395	-3.609	-3.533	-3.310

Table 2: Observed and expected Frequency distributions and metameter of geometric distribution

X_i	1	2	3	4	5	6
Obs.Freq . n K	205	66	12	03	03	01
				07		
Prob. $P(X_i)$	0.7	0.21	0.063	0.0189	0.0056	0.0017
Expected freq. $e_i = k * P(X_i)$	203	60.9	18.27	5.481	1.6443	0.4932
				7.6185		
$\dot{n}_k$	203.764	65.147	11.296	2.321	2.321	0.3676
Metameter φ n_k $\dot{\varphi}\ (n_k)$	-0.3529	-1.4932	-3.2453	-4.8275	-4.8275	-6.670

CamScanner

Taple 3: Observed and expected Bivariate Frequency distributions Trinomial distribution.

→ 4	0	1	2	3	4	5	Margina l Freq. of X2	Expected Marginal Freq.X2
	0(0.3)	1(1.12)	11(7.84)	26(27.4)	55(48.0)	27(33.6)	120	118.098
	0.(0.07)	2.(2.24)	12(11.7)	26(27.4)	24(24)	x	64	65.61
	0.(0.07)	2.(0.56)	7(5.88)	5.(6.68)	x	x	14	14.58
	0(0.03)	0(0.26)	2(0.98)	X	x	x	2	1.62
	0.(0.009)	0(0.03)	X	X	x	x	0	0.09 16.29
	0.(0.009)	X	X	X	x	x	0	0.002
$4.X_1$	0 05 05		32	57	79	27	200	200
$q.X_1$	0.486 5.670 6.156		26.46	64.74	72.03	33.61	200	

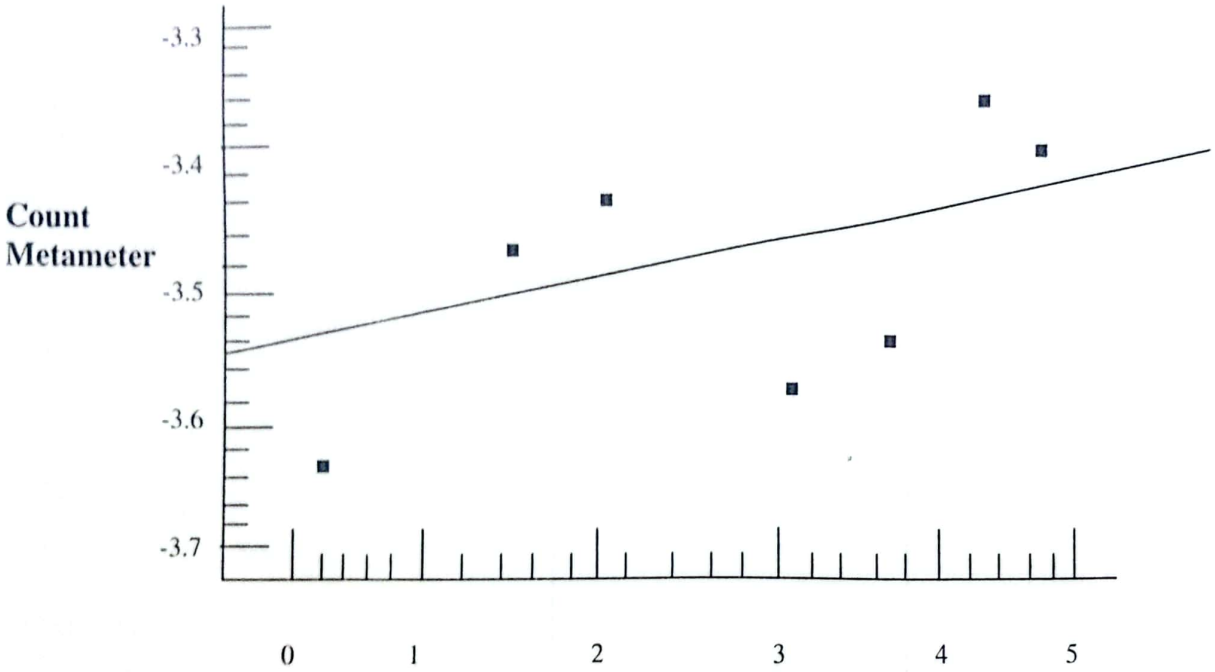
Table 4: Fitting of trivariate distribution.

$X_1 \rightarrow$ $X_2 \downarrow$	2	3	4	5	Marginal Freq.of X_2
0	12(8.99)	26(27.4)	55(48.02)	27(33.614)	120
1	14(14.01)	26(27.4)	24(24)	X	64
2	11(7.828)	5(6386)	X	X	16
Marginal freq.of X1	37	57	79	27	200

Exploratory plot for adequacy of simulated binomial random variate

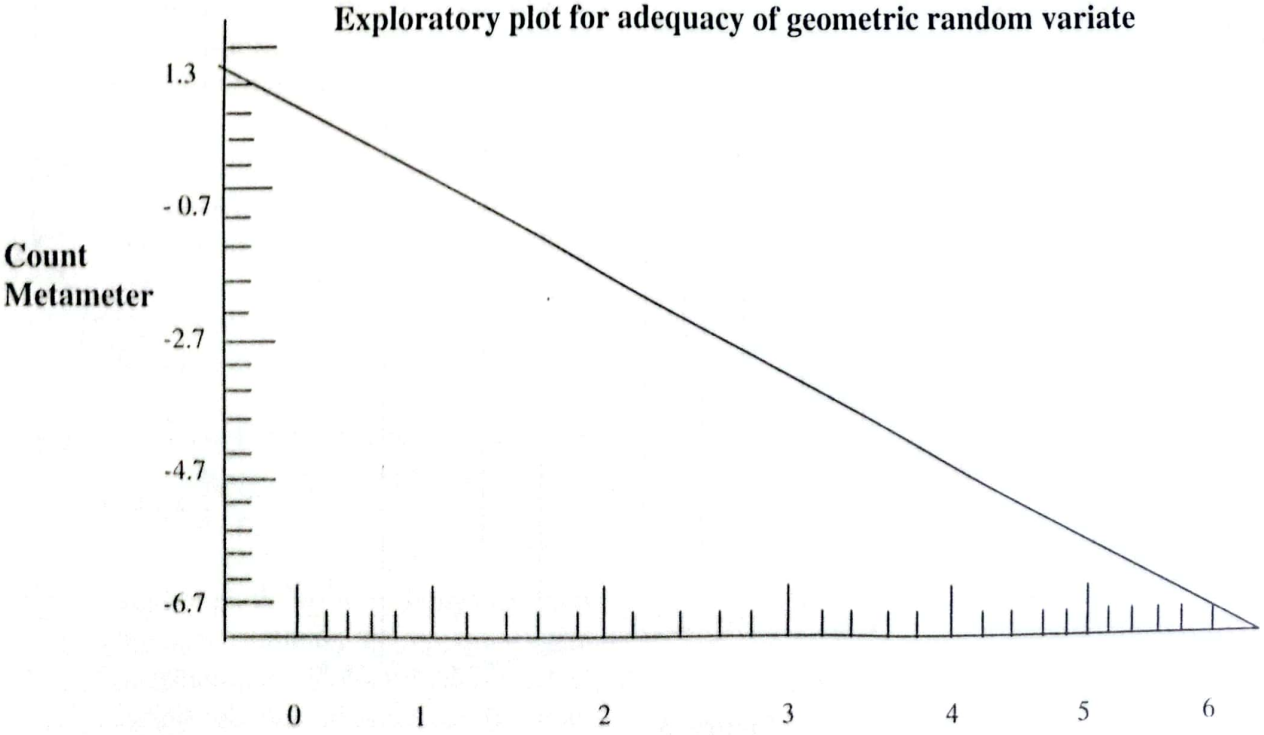
CamScanner

Exploratory plot for adequacy of simulated binomial random variate



Binomial Random Variate
Figure 1

Exploratory plot for adequacy of geometric random variate



Geometric Random Variate
Figure 2

CamScanner

Multiple line chart observed and expected cell frequency count metameter

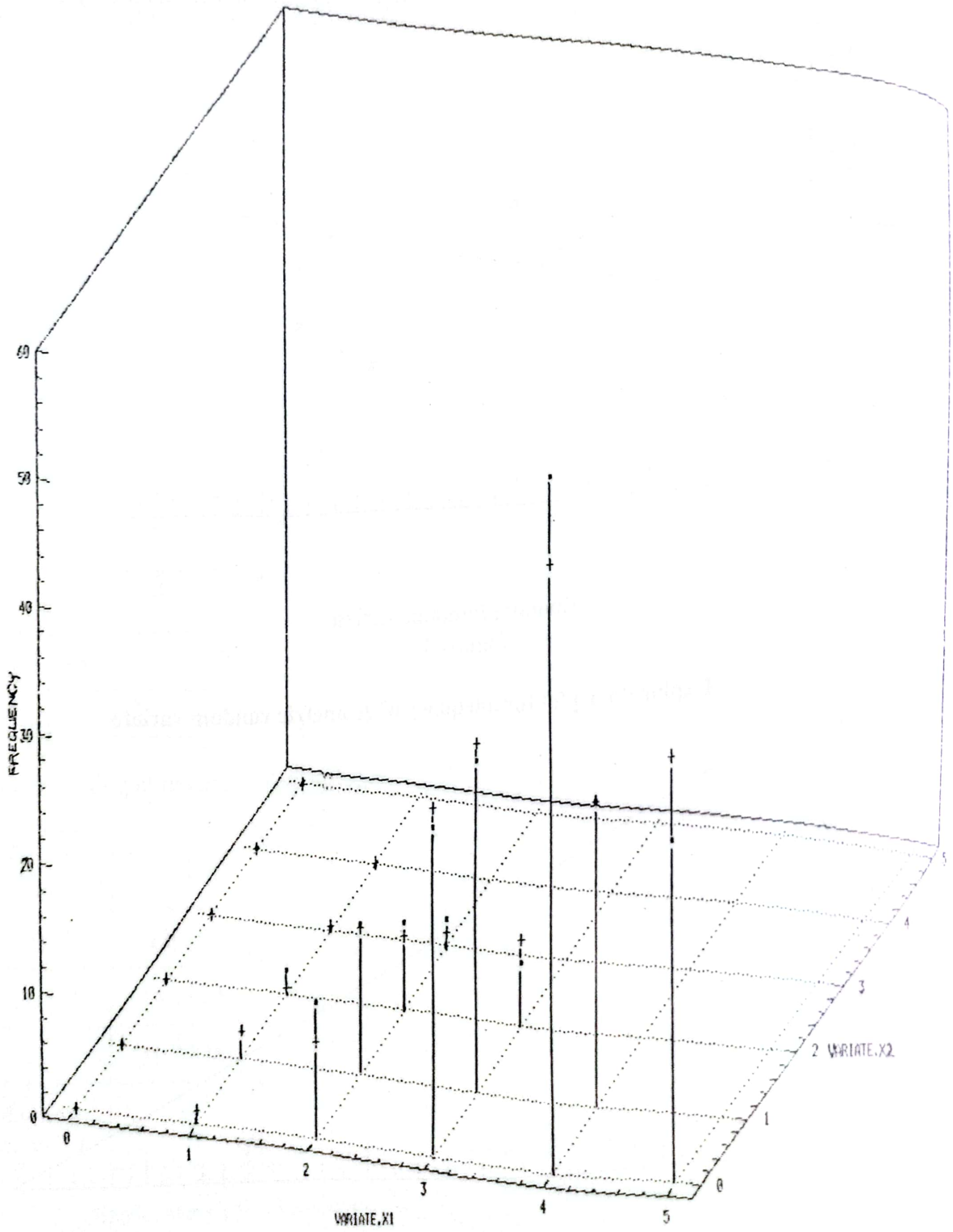


Figure 3

• Observed frequency

+Expected frequency

References

[1] Scott, W.F. and Lindley, D.V. (1984). New Cambridge Elementary Statistical Tables. Cambridge University Press, U.K.

[2] Hogg & Craig. (1978). Introduction to Mathematical Statistics, 4ed. Macmillan Publishing Co., Inc. New York.

[3] Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1985). Exploring Data Tables Trends and Shapes. John Wiely & Sons Inc. NY.

[4] Hormann, W.(1993a).The generation of binomial random variates. Journal of Statistical Computation & Simulation, 46,101-110.

[5] Hormann, W. (1993 b). The quality of non-uniform random numbers. DGOR/NSOR. Tagung. Amsterdam.

[6] Kachitrichyanukul, V. And Schmeiser, B.W. (1985). Computer generation of hypergeometric random variates. Journal of Statistical Computation and Simulation 22, 127-145.

[7] Morgan, B.J.T. (1984). Elements of Simulation. Chapman & Hall, London.

[8] Stadlober, E. (1989). Sampling from Poisson, binomial & hypergeometric distributions: Ratio of uniform as simple & fast alternative. Bericgte der Math. Stat. Sektion in der Forschungsgesellschaft Joanneum Nr.303. Graz.

[9] Shannon .R.E.(1975). System Simulation; The art and Science. Englewood Clifs, N.J.Prentice-Hall.

[10] STATGRAPHIC (1988). Statistical Graphic Corporation, U.S.A

CamScanner