ON THE SIMULATION OF SOME PARTICULAR DISCRETE DISTRIBUTIONS

Ion VĂDUVA, Mihăiţă DRĂGAN

University of Bucharest, Romania (vaduva@fmi.unibuc.ro, dragan_mihaita@yahoo.com)

DOI: 10.19062/1842-9238.2018.16.2.2

Abstract: *The paper summarizes the first several known discrete probability distributions which may describe the occurrence of a random number of events in various experiments, e.g in a reliability system, the occurrence of a random number of failures. Among these discrete distributions, the first mentioned are usual distributions such as: Poisson(X),\ > 0;Geometric(q),0 < q< l; Pascal(k,p),k £ N+, 0 < p < 1; Binomial(n,k,p),n,k £ N+,0 < p < 1. Then, some new discrete distributions are defined in terms of positive convergent series {an}, 1 < n < oo,an > 0. The paper presents methods of simulating the above mentioned distributions, which are either general, like the inverse method, or based on the rejection enveloping method. As enveloping distributions, either of the said distributions – i.e., Poisson and Geometric – or other less known distributions – such as the Zipf distribution or the Yule Distribution – are used. Comments related to testing these algorithms are finally presented.*

Keywords: *discrete probability distributions, Zipf distribution, Yule Distribution*

1. INTRODUCTION

Any discrete distribution, in the form $p_n = P(N = n), n = 1,2,...$ can describe the occurence of a random number of events. In reliability, some usual distributions of this type are used, basically, the following distributions [2,3,5,9], truncated on [l,oo), (i.e. $n > 1$).

a. Geometric distribution *Geo(p),* $O < p < 1$, definesd as

$$p_n = P(N = n) = q^n, n \text{ G } N^+, \tag{1.1}$$

b. Pascal distribution *Pas(p,k),* $0 < p < 1$, $k \in N^+$, defined as

$$p_n = P(N = n) = C_{n+k-1}^{k-1} \frac{p^k q^n}{1 - p^k}, n \in N^+, \tag{1.2}$$

c. *Poisson(λ),* $\lambda > 0$ distribution defined as

$$p_n = P(N = n) = \frac{1}{e^{\lambda} - 1} \frac{\lambda^n}{n!} e^{-\lambda}, n \in N^+ \tag{1.3}$$

d. Binomial distribution *Binomial(n,p),* $n \in N^+$, $0 < p < 1$ defined as

$$p_{\alpha} = P(N = \alpha) = \frac{1}{1 - q^n} C_n^{\alpha} p^{\alpha} q^{n-\alpha}, q = 1 - p, \alpha \in N^+. \tag{1.4}$$

Methods for simulating these distributions are presented in various papers (see [1,5,9]).

Note that any convergent series of positive terms a_n, $n \geq 1$, could define *a discrete distribution*.

If

$$\beta = \sum_{n=1}^{\infty} a_n, 0 < \beta < \infty, \tag{1.5}$$

then the probabilities of the discrete distributions derived from such series are

$$p_n = \frac{a_n}{\beta}. \tag{1.5'}$$

Some other known discrete distributions, applied in different circumstances are the following.

e. *The distribution of Euler [5]* defined as

$$p_n = \left(\frac{1}{n} - \ln\left(1 + \frac{1}{n}\right)\right)\frac{1}{\beta}, n \geq 1, \tag{1.6}$$

where

$$\beta = \gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln n\right) \tag{1.6'}$$

where γ is the *Euler's constant* (0 < γ < 1) [8].

f. The *distribution of Kemp* defined as [4]

$$p_n = -\frac{\alpha^n}{n \log 1 - \alpha}, n \geq 1, 0 < \alpha < 1, \tag{1.7}$$

This distribution is also called *logarithmic series distribution of the parameter p*, 0 < *p* < 1 and p_n is in equivalent form

$$p_n = \frac{a}{n}p^n, n = 1,2, \ldots a = -\frac{1}{\log(1-p)}. \tag{1.7'}$$

g. The *Zipf distribution* [1,5,6] of the parameter *a*, *a* > 1 defined as

$$p^n = \frac{1}{\zeta(a)n^a} n > 0, \tag{1.8}$$

where

$$\zeta(a) = \sum_{i=1}^{\infty}\frac{1}{i^a} \tag{1.8'}$$

is the *Riemann's function*. This distribution describes the occupied memory cells in the computer when the memory is dynamically allocated.

h. *The Yule(a)* distribution of the parameter *a*, *a* > 1, defined as [1,5,6]

$$p_n = P(N = n) = \frac{1}{c(a)}B(n,a), n \geq 1, c(a) = \sum_{n=1}^{\infty} B(n,a), \tag{1.9}$$

where *B(n,a)* is *Beta function* defined as

$$B(n,a) = \int_0^1 (1-u)^n u^a du, \tag{1.9'}$$

which is connected with function Γ(*p*) by the formula

$$B(n,a) = \frac{\Gamma(n)\Gamma(a)}{\Gamma(n+a)}. \tag{1.9''}$$

The function $\Gamma(p)$ is defined as

$$\Gamma(p) = \int_0^{\infty} u^{p-1} e^{-u} du, p \in R^+. \quad (1.9''')$$

Note that for $p \in N^+$, the function $\Gamma(p)$ is

$$\Gamma(p) = (p-1)!. \quad (1.9^{iv})$$

The simulation of these distributions is presented in several books and papers (see [1,4,5,7]) and they will be briefly described as such in a following section of this paper.

In [8] **several positive convergent series** $\{a_n\}_{n \in N^+}$ are found which could define such discrete distributions, as

$$P(N = n) = p_n = \frac{a_n}{\beta}, \sum_{n=1}^{\infty} p_n = 1.$$

The following is a list of positive convergent series collected from [8]:

$$a_n = \frac{n}{a^n}, a > 1, \sum_{n=1}^{\infty} a_n = \frac{a}{(1-a)^2} = \beta, \quad (1.10)$$

$$a_n = \frac{1}{n(n+p)}, p \in N^+, \sum_{n=1}^{\infty} a_n = \frac{1}{p} \sum_{k=1}^{p} \frac{1}{k} = \beta \quad (1.11)$$

$$a_n = \frac{1}{(n-1)! + n!}, \sum_{n=1}^{\infty} a_n = 1 = \beta \quad (1.12)$$

$$a_n = \frac{n}{(2n+1)!}, \sum_{n=1}^{\infty} a_n = \frac{1}{2} = \beta \quad (1.13)$$

$$a_n = \frac{n}{(n+1)!}, \sum_{n=1}^{\infty} a_n = 1 = \beta \quad (1.14)$$

$$a_n = \frac{n!}{(n+k)!}, \sum_{n=1}^{\infty} a_n = \frac{1}{(k-1)k!} = \beta \quad (1.15)$$

$$a_n = \frac{1}{n^2}, \sum_{n=1}^{\infty} \frac{1}{2^n} = \varsigma(2) \quad (1.16)$$

$$a_n = \frac{n^2}{n!}, \sum_{n=1}^{\infty} a_n = 2 + 2e = \beta \quad (1.17)$$

$$a_n = \frac{n^3}{n!}, \sum_{n=1}^{\infty} a_n = 5e = \beta \quad (1.18)$$

$$a_n = \frac{(p+1)\ldots(p+n)}{(q+1)\ldots(q+n)}, q - p > 1, \beta = \frac{p+1}{q-p-1} \quad (1.19)$$

One aim of this paper is to present methods for simulating the distributions defined by (1.10)…(1.19).

2. THE INVERSE METHOD

Any probability distribution can be simulated by a general method, *the inverse method* [1,5,6,7,9]. If $F(x) = P(X < x)$ is the c*umulative distribution function (cdf)* of a random variable *X*, then a sampling value of *X* is simulated by the formula $X = F^{-1}(U)$, where *U* is a random number, uniformly distributed over (0,1). (See [1,6,7,9,10]). This induces the following algorithm:

Algorithm INV
begin
generate U an uniform random number over (0,1);
take $F^{-1}(U)$*, (where* F^{-1} *is the inverse of function F(x)=P(X < x));*
end.

The method can be used if there is an easy way to calculate the inverse function $F^{-1}(U)$. In the discrete case, where the function *F* is a "step" function, the jumps of the function are in the points 1,2,…. Thus, the distinct values of *F(x)* are *F(i)* defined as

$$F(i) = \begin{cases} 0, if\ x < 0 \\ \sum_{\alpha=0}^{i} p_\alpha, if\ i \le x < i+1, i = 1,2,\dots \end{cases} \quad (2.1)$$

To simulate a random sampling value *i*, we must calculate $F^{-1}(U)$ using *F* in the formula (2.1) as a step function. In other words, we must *search* the index *i*, such as $F(i) \le U < F(i+1)$. There are various possibilities to search *i*. One could be *binary search*. Here we use a simpler (but not faster!) procedure, based on dividing the interval (0,1) in *five* intervals, namely

$(0,1) = (0,0.25) \cup [0.25,0.5) \cup [0.5,0.75) \cup [0.75,0.95) \cup [0.95, \infty)$.

The algorithm uses a table of distinct values of $F(i), 1 \le i \le i_0$, such as $F(I_0) = 0.95$ (i.e. $F(I_0)$ is large enough). The values $F(i)$ are calculated as follows:

$$F(i) = \begin{cases} p_1\ if\ i = 1 \\ p_1 + p_2\ if\ i = 2 \\ \dots\dots\dots \\ p_1 + p_2 + \dots + p_k\ if\ i = k \end{cases} \quad (2.1')$$

(The means by which *k* can be determined is explained later; it is the I_0 index below).

Let us select the indexes I_1, I_2, I_3, I_0 as follows:

$$\begin{cases} F(I_1) \le 0.25 < F(I_1+1); & F(I_2 \le 0.5 \le F(I_2+1); \\ F(I3) \le 0.75 < F(I_3+1); & F(I_0) \le 0.95 < F(I_0+1). \end{cases} \quad (2.2)$$

The detailed algorithm **INV** is the following

Preparatory step; Calculate F(1),…F(I_0), determine I_1, I_2, I_3, I_0.

1. **generate** U **Uniform on** (0,1).
2. **if** U ≤ 0.25 **then begin**

i = I_1; **while** U ≤ F(i) **do** i := i – 1 **end else if** U ≤ 0.5 **then**
begin i = I_2; **while** U ≤ F(i) **do** i := i – 1 **end**
else if U ≤ 0.75 **then begin** i = I_3; **while** U ≤ F(i) **do** i := i – 1 **end**
else if U ≤ F(I_0) **then begin** i := I_0; **while** U ≤ F(i) **do** i := i – 1 **end**
else begin i := I_0; **while** U > F(i) **do begin** i := i + 1; F(i) := F(i) + p_i
end; end;
deliver i.

(i.e. "i" is the generated sampling value). As $F(i), 1 \leq i \leq I_0$ are calculated only once and if I_0 is small, then the algorithm is fast for generating a sampling value *i*. But sometimes I_0 may not be small at all, and then the algorithm will be slow.

The detailed algorithm, described in steps **1.** and **2.** can be adapted and applied to each of the distributions (1.10-1.19).

3. THE ACCEPTANCE-REJECTION METHOD

There are various versions of this method (see [1,4,5,6,7,9,10]). Here, we will be using the rejection method based on **enveloping** the frequency function $f(n) = p_n$ of the distribution with another frequency function $h(n) = q_n$, which can be simulated. Let us assume that there is a constant $\alpha > 1$ such as $\frac{f(n)}{h(n)} < \alpha, n = 1,2, \ldots$.

The formal Theorem is the following: if X is a random variable with frequency function p_n (to be simulated) and if Y is another random variable (which can be simulated) whose frequency function is h_n, and if there is a constant $\alpha, 1 < \alpha < \infty$ such as

$$\frac{p_n}{h_n} \leq \alpha,$$

and if U is an uniform (0,1) random number independent of Y, then if

$$0 < U \leq \frac{p_n(Y)}{\alpha h_n(Y)},$$

the simulated value of X is X = Y.

The general simulation algorithm is:

Algorithm REJ
repeat
simulate a random variate U uniform (0,1);
simulate j a random variate with the frequency function h(n);
until $U \leq \frac{f(j)}{\alpha h(j)}$;
deliver i = j;

The value *i* is the simulated sampling value of *f(n)*.

The acceptance probability of the algorithm is $p_a = \frac{1}{\alpha}$ and if it is large, then the algorithm is fast. The function *h* is *the enveloping function*. To build up the algorithm **REJ**, it is important to find **a good** enveloping function *h* such a way as the **acceptance probability** which is large.

Discussions on simulating by rejection procedure REJ any of distributions (1.10)-(1.19) will be based on the following idea: the distribution *h(n)* could be a distribution which is decreasing with *n*, such as can happen in cases of convergent series with positive terms.

This suggests that sometimes (but not always), a candidate for *h(n)* could be the **geometric distribution** *Geom(p)*, $0 < p < 1, q = 1 - p$, for which

$$p_n = pq^n, 0 < p < 1, q = 1 - p, n = 0,1,2 \ldots \quad (3.1)$$

truncated to $n \geq 1$.Therefore, in this case, the distribution *h(n)* has probabilities

$$h_n = \frac{p}{1-q} q^n = q^n, n = 1,2, \ldots \quad (3.2)$$

An enveloping candidate could be also *Poisson* or any other selected distribution.

First we have to specify how to simulate the truncated distribution *Geom(p)*, $n \geq 1$.

In [1,5,9] two procedures to simulate this distribution are presented. Note that this distribution is related to **Bernoulli triles**.

A Bernoulli trile is an experiment on an event with *constant probability p* which, when it occurs, we say that is a *success* and when it does not occur, we say that is a *failure*. The *number of failures N until a success occurs* is a random variable having distribution *Geom(p)*. Therefore, this can be simulated as:

Algorithm COUNT FAILURES

1. *Read p*, $0 < p < 1; j := 0;$

2. Repeat

Generate U uniform (0, 1); if U ≥ p then $j: + j + 1$

until U < p.

The value *j* is the simulated value of *Geom(p)*.

We can also use *the inverse method* to simulate *Geom(p)*. The cdf in this case is

$$F(n) = P(N < n) = \sum_{i=0}^{n-1} pq^i = p\frac{1-q^n}{1-q} = 1 - q^n. \qquad (3.3)$$

The inverse method gives

$$j = \left[\frac{U}{q}\right] \qquad (3.3')$$

where [t] means the integer closest to real number t.

Since the values of *j* must be positive, we have to *reject* the value of j = 0, i.e. the algorithm is:

repeat

generate j from Geom(p)

until *j* > 0.

The value *j* is the simulated value of the **truncated** *Geom(p)*.

In order to build up the algorithm **REJ** for all discrete distributions in the form (1.10)-(1.19), it is enough, in each case, to specify the *possible envelope distribution and then to determine the constant* α in the algorithm **REJ**. For instance, to determine q of the enveloping *Geom*(p), we find first the maximum value of a_n and if this is $\beta = a_m$, then select q in the form $q = \beta$, β will be a normalizing number.

3.1 Simulation of the distribution defined by (1.10).

Method 1. Here is where we try to determine the *geometric distribution* as envelope.

The maximum of a_n is determined as the maximum of the function

$$f(x)_{max} = \max\left(\frac{1}{\beta}\frac{x}{a^x}\right), \qquad x \geq 0.$$

i.e. the maximum of

$$f(x) = xa^{-x}, \qquad a > 1.$$

After some calculations, it results that the maximum point of this function is

$$x_0 = \sqrt{\frac{a}{\ln(a)}} > 1, \qquad \ln(a) > 0,$$

and hence

$(a_n)_{max} = \frac{(a-1)^2}{a}\frac{x_0}{a^{x_0}}$, and this gives

$$q = \frac{(a-1)^2}{a}\frac{x_0}{a^{x_0}}. \quad (3.4)$$

To determine α, consider the ratio

$$r_n = \frac{a_n}{q^n}, \qquad i.e. r(x) = \frac{f(x)}{q^x}$$

which, in a similar manner has the same maximum point x_0 , and after some calculation we finally obtain

$$\alpha = \frac{a^{x_0}+1}{a^{x_0}+1-(a-1)^2 x_0} > 1. \quad (3.4')$$

Now, the construction of the algorithm **REJ** is terminated.

Method 2. An alternative method of simulating this distribution is to use *the inverse method* of the *equivalent* distribution directly:

$$f(n) = \frac{b^3}{(b-1)^2} n b^{n-1} = K^* n b^{n-1}, \qquad b = \frac{1}{a} < 1, K^* = \frac{b^3}{(b-1)^2}.$$

The cdf is therefore

$$F(n) = K^* \sum_{i=1}^{n} i b^{i-1} = K^* \left(\sum_{i=1}^{n} b^n\right)' = K^* \left(\frac{b^{n+1}-b}{b-1}\right)' = K^* \frac{(n+1)b^n}{b-1}$$

where *the derivative ()'* is calculated with respect to n. To apply the inverse method, we have to solve in n (numerically!) the equation

$$F(n) = U, i.e., K^* \frac{(n+1)b^n}{b-1} = U. \quad (3.4'')$$

where U is an uniform random number over (0,1). If n_0 is the solution of (3.4'') then the simulated value is n = int(n_0).

Method 3. Let us use as enveloping distribution the *Kemp distribution* i.e.

$$h_n = -\frac{1}{\log(1-p)}\frac{p^n}{n}, 0 < p < 1.$$

Then the ratio r_n becomes

$$r_n = -\frac{(a-1)^2 \log(1-p)}{a}\frac{n^2}{b^n}, b = ap.$$

If b>4, it is shown by induction that

$$\frac{n^2}{b^n} < 4$$

Therefore, when p is selected such as ap > 4 then α > 1 and the algorithm **REJ** is obvious. With respect to p_a, it seems that **method 3** is preferable.

3.2 Simulation of the distribution defined by (1.11)

Method 1. Note that the sequence is

$$a_n = \frac{1}{\beta}\frac{1}{n(n+p)}, p_n = \frac{a_n}{\beta}, \beta = \frac{1}{p}\sum_{k=1}^{p}\frac{1}{k} = \frac{1}{p}K^*.$$

Let us choose this time as *enveloping distribution* a Zipf(2) distribution [1,5] defined as

$$q_n = \frac{1}{\zeta(2)n^2}, \qquad \zeta(a) = \sum_{n=1}^{\infty}\frac{1}{n^2}. \quad (3.5)$$

Consider the ratio

$$r(n) = \frac{p_n}{q_n} = \frac{\zeta(2)}{\beta}\frac{n^2}{n(n+p)}.$$

After some simple calculations we have

$$r_n \le \frac{\zeta(2)}{\sum_{k=1}^{p}\frac{1}{k}} = \alpha > 1 \qquad (3.5')$$

and the construction of the algorithm **REJ** is finished. Simulation of the Zipf distribution is found in [1,5] and is presented in the last section. There is a version of this distribution [1] which is defined for n = 1,2,…$K^* < \infty$ (i.e. a finite series!), referring to a finite population of size K^*. Comments on this, will me made in the last section of the paper.

Method 2. Let us take as enveloping distribution that given by (1.10). Therefore we have

$$p_n = \frac{p}{\sum_{k=1}^{p}\frac{1}{k}}\frac{1}{n(n+p)}, h_n = \frac{(a-1)^2}{a}\frac{n}{a^n}.$$

then the ratio is

$$r_n = \frac{pa}{(a-1)^2\sum_{k=1}^{p}\frac{1}{k}}\frac{a^n}{n^2(n+p)} = \frac{pa}{(a-1)^2 S}\frac{a^n}{n^2(n+p)}, \qquad S = \sum_{k=1}^{p}\frac{1}{k}$$

Note that ratio

$$R_n = \frac{a^n}{n^2(n+p)} < \frac{a}{p+1}, a > 1, p \ge 1,$$

proved by induction. Therefore

$$r_n < \alpha < \frac{pa}{(a-1)^2 S}\frac{a}{p+1},$$

Which if $a \ge p+1$ gives α > 1, and the algorithm **REJ** is obvious. To decide which of these method is preferable, it is necessary to numerically compare the p_a of the two methods.

3.3 Simulation of the distribution defined by (1.12)

Method 1. For the sequence a_n , we have

$$a_n = p_n = f(n) = \frac{1}{(n-1)!\,(n+1)}, \beta = 1,$$

we select the enveloping distribution h(n) as a *Poisson(1)* i.e.

$$q_n = \frac{1}{(n-1)!}, \sum_{i=1}^{\infty} q_n = e, h(n) = \frac{1}{e}\frac{1}{(n-1)!} \qquad (3.6)$$

The ratio

$$r(n) = \frac{f(n)}{h(n)} = \frac{e(n-1)!}{(n-1)!\,(n+1)} \le \frac{e}{2} = \alpha, \qquad \alpha > 1. \qquad (3.6')$$

Since elements f(n), h(n), α are specified, the algorithm **REJ** is obvious.

Method 2. If we select as enveloping distribution that given by (1.11), then we have

$$r_n = \frac{n(n+p)}{(n-1)!\,(n+1)}\frac{1}{\beta}$$

and

$$r_n = \frac{n^2(n+p)}{(n+1)n(n-1)!\,\beta} \le \frac{p+1}{\beta} = \alpha, \qquad \alpha = \frac{(p+1)\sum_{k=1}^{p}\frac{1}{k}}{p} > 1$$

therefore, the algorithm **REJ** is defined.

Method 3. Since $p_n \le \frac{1}{2}$, we can take as envelope the *Geo*(p), p = $\frac{1}{2}$ and

$$r_n = \frac{2.2^n}{(n-1)!\,(n+1)} \le \frac{4e}{2} = 2e^2 > 1$$

and again, the required algorithm is ready. Note that **method 1** is the best of the three methods, since in that case α is close to one.

3.4 Simulation of the distribution defined by (1.13)

Method 1. The sequence

$$a_n = \frac{n}{(2n+)!}, \beta = \frac{1}{2}, gives\ p_n = \frac{2n}{(2n+1)!}$$

If we select as enveloping distribution the *Poisson(1)* distribution in the form

$$h_n = \frac{1}{e}\frac{1}{(n-1)!}, n \ge 1, \tag{3.7}$$

we obtain

$$r_n = \frac{p_n}{h_n} = \frac{2en(n-1)!}{(2n+1)!} \le \frac{2e}{2} = \alpha = e. \tag{3.7'}$$

The algorithm **REJ** is obvious in this case.

Method 2. Let us take as envelope the distribution (1.12). In this case, we have

$$r_n = \frac{2n(n+1)(n-1)!}{(2n+1)!} = \frac{2(n+1)!}{(2n+1)!} \le \frac{2(2n+1)!}{(2n+1)!} = 2 = \alpha \tag{3.3''}$$

and algorithm **REJ** is obvious. This method is better than **method 1**, since p_a is larger.

3.5 Simulation of the distribution derived from (1.4)

Method 1. In this case

$$a_n = p_n = \frac{n}{(n+1)!}.$$

We select as an enveloping distribution the *Poisson*(1), i.e.

$$q_n = \frac{1}{(e-1)n!}. \tag{3.8}$$

The ratio r_n is

$$r_n = \frac{p_n}{q_n} = \frac{n(e-1)}{n+1} \le e - 1 = \alpha \tag{3.8'}$$

and elements of the algorithm **REJ** are determined.

Method 2. Let us choose as enveloping distribution the *Kemp distribution.* Then the ratio becomes

$$r_n = -\log(1-p)\frac{n^2}{(n+1)!\,p^n} \le -\log(1-p)\frac{1}{2p} = \alpha, 0 < p < 1. \tag{3.8''}$$

If we choose p such as $e^{2p} > 1 - p$, then α > 1 and algorithm **REJ** is defined. With respect to p_a, **method 1** is preferable.

3.6 Simulation of the distribution derived from (1.15)

In this case, note that

$$a_n = \frac{1}{\beta}\frac{n!}{(n+k)!}, \beta = \frac{1}{(k-1)k!}, a_n = \frac{n!\,k!\,(k-1)}{(n+k)!}$$

We choose as enveloping distribution *the Yule*(k) *distribution* in the form

$$h_n = \frac{B(n,k)}{c(k)}\,, c(k) = \sum_{n=1}^{\infty} B(n,k). \tag{3.9}$$

The ratio r_n is

$$r_n = \frac{n!\,k!\,(k-1)}{(n+k)!}\,\frac{(n+k-1)!\,c(k)}{(n-1)!\,(k-1)!} = \frac{nk(k-1)c(k)}{n+k} \le \frac{k(k-1)c(k)}{k+1}.$$

Therefore

$$\alpha = \frac{k(k-1)c(k)}{k+1} > 1\,. \tag{3.9'}$$

All elements of the algorithm **REJ** are defined. The probability p_a can be calculated numerically.

3.7 Simulation of the distribution derived from (1.16)

Method 1 (known). This is the *Zipf(a) Distribution*, defined in its general form as

$$p_n = \frac{1}{\zeta(a)}\,\frac{1}{n^a}\,, a > 1, \zeta(a) = \sum_{i=1}^{\infty} \frac{1}{i^a}\,, \tag{3.10}$$

where the ζ(a) is the ζ *Riemann function*. The formula (3.10) shows that ζ(2) ≤ 2, (See [1,8]). An algorithm for simulation of the random variable X as Zipf(a) is presented in [1,4]. It uses as enveloping distribution the distribution of a random variable Y such as

$$q_i = P(Y = i) = \frac{1}{(i+1)^a}[(1 + \left(1 + \frac{1}{i}\right)^{a-1} - 1], i \in N+, i > 1$$

For which the cdf is

$$H(i) = 1 - \frac{1}{i^{a-1}},$$

And the inverse method gives

$$Y = int(U^{\frac{-1}{a-1}}), \tag{3.10'}$$

(where *int* denotes "integer part"). Note that ratio

$$r_i = \frac{p_i}{q_i} < \frac{p_1}{q_1} = \frac{2^{a-1}}{\zeta(a)(2^{a-1}-1)} = \alpha > 1. \tag{3.10''}$$

Therefore, the algorithm is

Step1: Take b = 2^{a-1};

Step2: **repeat**

Generate U,V uniforms on (0,1), independent;

Take Y = int $(U^{\frac{1}{a-1}})$, *T*=$(1 + \frac{1}{Y})^{a-1}$;

Until VY $\frac{T-1}{b-1} \le \frac{T}{b}$;

Deliver X = Y.

X is the simulated value. The probability $p_a$can be easy approximated.

Method 2. Let us take as enveloping distribution, *the distribution derived from (1.10)* with the same a. Then we have

$$r_n = \frac{a}{(a-1)^2\zeta(a)}\,\frac{a^n}{na^n} < \frac{a}{(a-1)^2\zeta(a)} = \alpha, \alpha > 1 \tag{3.10'''}$$

and the algorithm **REJ** is terminated. To decide which of these methods is preferable, it is necessary to compare the p_a probabilities. The second method appears to be the best.

3.8 Simulation of the distribution derived from (1.17)

Method 1. The distribution is

$$p_n = \frac{1}{\beta}\frac{n^2}{n!}, \beta = 2 + 2e = 2(1 + e).$$

We take as enveloping distribution the Zipf(2) distribution defined as

$$h_n = \frac{1}{\zeta(2)n^2}, \zeta(2) = \sum_{i=1}^{\infty}\frac{1}{i^2}, \quad (3.11)$$

where $\zeta(2)$ is the Riemann function of the argument 2. Therefore

$$r_n = \frac{\zeta(2)}{2(1+e)}\frac{n^4}{n!}.$$

It is shown by induction that

$$\frac{n^4}{n!} \leq 4^2.$$

Therefore

$$r_n \leq \frac{\zeta(2)}{(1+e)} = \alpha > 1. \quad (3.11')$$

The algorithm **REJ** is specified

Method 2. Let us take as enveloping distribution the Kemp distribution. Then, the ratio $r_n$becomes

$$r_n = -\frac{\log(1-p)}{2(1+e)}\frac{n^3}{n.n!p^n} \leq -\frac{\log(1-p)}{2p(1+e)} = \alpha. \quad (3.11'')$$

As parameter p is free, we can choose it as follows:

$$(1-p)e^{2p(1+e)} < 1, and\ then\ \alpha > 1. \quad (3.11''')$$

Thus, algorithm **REJ** is terminated. To select the best method, the probabilities $p_a$must be estimated numerically.

3.9 Simulation of the distribution derived from (1.18)

Method 1. The distribution is

$$p_n = \frac{1}{5e}\frac{n^3}{n!}.$$

In this case we take again as enveloping distribution the *Zipf(2) distribution*, i.e.

$$h_n = \frac{1}{\zeta(2)n^2}. \quad (3.12)$$

The ratio r_n in this case is $r_n = \frac{\zeta(2)}{5e}\frac{n^5}{n!}$.

Here, again by induction, it is shown that

$$\frac{n^5}{n!} < 64,$$

and finally $\alpha = \frac{\zeta(2).64}{5e} > 1.$ (3.12')

The algorithm **REJ** is obvious.

Method 2. Let us choose as enveloping distribution the Kemp distribution $0 < p < 1$,

i.e. $h_n = -\frac{1}{\log(1-p)}\frac{p^n}{n}$. (3.12'')

The ratio r_n is

$$r_n = -\frac{\log(1-p)}{5e}\frac{n^4}{n!p^n} \leq -\frac{\log(1-p)}{p.5e} = \alpha = -\frac{\log(1-p)}{loge^{5ep}}. \quad (3.12''')$$

If we choose p as

$(1-p)e^{5ep} < 1,$ which can be done, then $\alpha > 1$, algorithm **REJ** is ready.

3.10 Simulation of the distribution derived from (1.19)

Method 1. In this case we have

$$p_n = \frac{q-p-1}{p+1}\frac{(p+1)\dots(p+n)}{(q+1)\dots(q+n)}, q-p>1. \quad (3.13)$$

Let us take as enveloping distribution (1.10)

$$h_n = \frac{(a-1)^2}{a}\frac{n}{a^n}. \quad (3.13')$$

Now the ratio $r_n = \frac{p_n}{h_n}$ is $r_n = \frac{(p+n)!}{(q+n)!}\frac{q!}{p!}\frac{a}{(a-1)^2}\frac{a^n}{n}$.

Since the function $\frac{a^x}{x} < 1$ for $x > int\left(\frac{1}{\log(a)+1}\right) = k$, we have

$$r_n \le \frac{(p+k)!}{(q+k)!}\frac{q!}{p!}\frac{a^2}{(a-1)^2}\frac{q-p-1}{p+1}.$$

If we now choose α as

$$\frac{(p+k)!}{(q+k)!(q-p-1)}\frac{a^2}{(a-1)^2} = \alpha, \quad (3.13'')$$

then the ratio becomes $r_n \le \alpha, \alpha \ge 1$, and the algorithm **REJ** is ready.

Method 2. Let us take as enveloping distribution the distribution of Kemp of the paramtere β, $0 < \beta < 1$, then we have

$$r_n = \frac{(p+n)!}{(q+n)!}\frac{q!}{p!}\frac{q-p-1}{p+q}\left[-\log(1-\beta)\frac{n}{\beta^n}\right].$$

Note that $\beta = e^{-\lambda}, \lambda > 0$, and then

$$\frac{n}{\beta^n} < \frac{n^e}{e^{-\lambda n}} < \frac{e^n}{e^{-\lambda n}}.$$

Finally, one obtain $r_n < \frac{p+1}{q+1}\frac{q-p-1}{p+q}[-\log(1-\beta)]e^{1+\lambda} = \alpha > 1.$ (3.13''')

i.e. the algorithm **REJ** is specified. Here again the probabilities $p_a$will show which method is preferable.

4. ADDENDA: SIMULATION OF USED DISTRIBUTIONS

The simulation of discrete distributions mentioned in the formulas (1.6)-(1.9) will be presented in the following.

4.1 Simulation of logarithmic series of the parameter p

This distribution is $p_n = -\log(1-p)\frac{p^n}{n}, n \ge 1, 0 < p < 1.$ (4.1)

Method 1. One method for simulating this distribution consists in the fact that the random variable X having this distribution is a mixture (see [5]) of the random variable Y with the cdf

$$F(y) = \frac{\log(1-y)}{\log(1-p)}, 0 \le y \le p. \quad (4.2)$$

with the *Geometric*(y) distribution. Therefore, the algorithm is *Generate a random variate y by the inverse method,* i.e solve the equation

$$\frac{\log(1-y)}{\log(1-\alpha)} = U; \quad (4.2')$$

Generate X as Geometric(y).

Deliver X.

In [5], the inverse method for simulating X is presented, (i.e. the solution of (4.2')).

Method 2. Let us take as enveloping distribution the one deriving from (1.10).

Then,

$$r_n = -\log(1-p)\frac{p^n}{n}\frac{a}{(a-1)^2}\frac{a^n}{n} = -\log(1-p)\frac{a}{(a-1)^2}\frac{(pa)^n}{n^2} \le -\log(1-p)\ \frac{a}{(a-1)^2}pa,$$

$if\ pa < 1$

In this case $r_n \le\ -\log(1-p)\frac{a^2p}{(a-1)^2} = \alpha,$

which gives $\alpha > 1\ if\ 1 < a < 2$ and the algorithm **REJ** is defined. It seems difficult to compare these methods, if not by means of computer tests.

4.2 Simulation of $Zipf(a), a > 1$ distribution

Method 1. In this case

$$p_n = \frac{1}{\varsigma(a)n^a}, n \ge 1, \varsigma(a) = \sum_{n=1}^{\infty}\frac{1}{n^a}, \quad (4.3)$$

where $\varsigma(a)$ is the *Riemann function.* In [1,5] a *rejection method* which uses the enveloping distribution is presented:

$$q_n = p(Y = n) = \frac{1}{(n+1)^{a-1}}\left[\left(1+\frac{1}{n}\right)^{a-1} - 1\right], n \ge 1. \quad (4.4)$$

By calculating the ratio $r_n = \frac{p_n}{q_n}$, it results that $r_n = \frac{p_n}{q_n} \le \frac{p_1}{q_1} = \frac{2^{a-1}}{\varsigma(a)(2^{a-1}-1)} = \alpha > 1.$ (4.5)

In [1] it is shown that

$$\alpha \le \frac{12}{\pi^2}\ if\ a \ge 2\ and\ \alpha \le \frac{2}{\log(2)}, if\ 1 < a < 2\,. \quad (4.5')$$

There are some remarks to be made regarding this distribution.

(1). If $1 \le n \le K^* < \infty$, it is used to represent random events, such as number of occupied cells of a computer memory of size $K*$, when memory is dynamically allocated;.

(2). For a finite $K*$ this distribution describe the random occurrence of words in a text of a given length (natural language).

Method 2, (New method). Let us select as enveloping function the h_n as the $Kemp$ distribution. Then

$$h_n = -\frac{1}{\log(1-p)}\frac{p^n}{n}, 0 < p < 1\ ,$$

and hence

$$r_n = \frac{p_n}{h_n} = -\frac{\log(1-p)}{\varsigma(a)}\frac{n}{n^a p^n} \le -\frac{\log(1-p)}{p} = \alpha. \quad (4.5'')$$

We can choose the parameter p such as $\alpha > 1$ and the algorithm **REJ** is obvious. A more relevant comparison between these methods could be done by means of computer tests.

4.3 Simulation of Euler distribution

Method 1.(Known). This distribution is

$$p_n = \left[\frac{1}{n} - \log\left(1+\frac{1}{n}\right)\frac{1}{\gamma}\right], \gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \log(n)\right), \quad (4.6)$$

where γ is *the constant of Euler*. In this case, we are using a rejection method based on enveloping p_n, with the *distribution of Logarithmic series of parameter p, 0¡p¡1*. The ratio is

$$r_n = -\frac{1}{\gamma\log(1-p)}\left(\frac{1}{n} - \log\left(1+\frac{1}{n}\right)\frac{n}{p^n} = -\frac{1}{\gamma\log(1-p)}\right.$$

$$= 1 - \frac{\log\left(1+\frac{1}{n}\right)^n}{p^n} \le -\frac{1}{\log(1-p)}\frac{1-\log(2)}{p} = \alpha. \quad (4.6')$$

If we choose $p, 0 < p < 1$ such as

$$e > 2 + \left(\frac{1}{1-p}\right)^p$$

then $\alpha > 1$. The elements of the algorithm **REJ** are defined.

Method 2.(New) Let us take as enveloping distribution the *Geometric(q)*, where $q = \max\left(\frac{1}{n} - \log\left(1 + \frac{1}{n}\right) = \log\left(\frac{e}{2}\right) = q < 1.\right.$

Since $= \log(2)$, we have

$$r_n \leq \frac{\log\left(\frac{e}{2}\right)}{\log(2)\,\gamma\left(\log\left(\frac{e}{2}\right)\right)} = \frac{1}{\log(2)\,\gamma} = \alpha > 1,$$

and **REJ** is defined. Here again, the comparison of methods could be done via computer tests.

4.4 Simulation of $Yule(a)$ distribution

The simulation is based on the following judgment: *The Yule(a) distribution is the mixture of the Geometric(p) distribution with*

$p = e^{-\frac{Y}{a-1}}$

and Exp(1) distribution of Y. This results in the following algorithm:

1. *Generate E and Exp(1) random variate. (i.e. Generate U uniform (0,1) and take E=-log(u)), U>0.*

2. *Generate $E^* \mapsto Exp(1)$, independent from E;*

3. *Calculate*

$$X = \text{int}\left\{\frac{E}{\log\left(1-e^{-\frac{E^*}{a-1}}\right)}\right\} + 1.$$

The X variable is the required Yule(a) variable.

In [1], it is specified that the Yule distribution is a better approximation of word frequencies (in a natural language) than the Zipf distribution.

Comments. Computer tests were not performed yet. They could be performed following the hints in [10]. This could make a good exercise for an M.Sc. student. Such an exercise could be could be useful for comparing various methods of simulation for each distribution. The **inverse** algorithms must be first considered to assess the performance degree of these methods.

REFERENCES

[1] Luc Devroye, (1986).*Non Uniform Random Variate Generation*, Springer Verlag, New York, Berlin.

[2] I. B. Gerstbakh, (1989). *Statistical Reliability Theory*, Marcel Dekker, Inc., New York, London.

[3] N. L. Johnson and S.Kotz, (1972). Distributions in Statistics: Discrete univariate distributions,John Wiley and Sons, New York, London.

[4] A.W. Kemp, (1981). *Efficient generation of logarithmically distributed pseudo-random variables*, Appl. Statistics, 30(3), p.249-253.

[5] J.T. Byron Morgan (1984).*Elements of Simulation*, Chapman and Hall, London, New York.

[6] B. D. Ripley (1987). *Stochastic Simulation*, John Wiley and Sons, New York.

[7] Sheldom R.M. (1997). *Simulation*. Second Edition, Academic Press, San Diego, New York, London.

[8] Ghe. Siret (1985). *Calcul Diferential si Integral*, Vol.1 Editura Stiintica si Enciclopedica. Bucuresti.

[9] I. Vaduva, (1977). *Simulation Models with Computers*, (Romanian), Ed. Tehnica, Bucharest.

[10] I. Vaduva, (2011).*On Simulation of some Mixed Life Distributions*. Analele Univ. Bucuresti, Seria Informatica,Anul 2011, p.10-19.