SONDERDRUCK AUS

OPERATIONS RESEARCH VERFAHREN METHODS OF OPERATIONS RESEARCH

XXIX

An Application of Record Values to Stochastic Simulation

Dietmar Pfeifer, Aachen

Abstract:

Using record values rather than record times, a new method for testing the independence of random number generators with continuous cumulative distribution function (c.d.f.) is proposed as well as another algorithm for generating Poisson-distributed random variables (r.v.'s) from any continuous distribution.

Record values were originally introduced by K.N. Chandler in 1952, inspired by "the frequency with which record weather conditions are reported in the newspapers" ([1]).

If $\{X_k\}_{k=1}^{\infty}$ is a sequence of real r.v.'s on a probability space $(\Omega, \mathfrak{A}, P)$, any observation of this sequence which is strictly greater or less than all previous ones is called a record value. The indices at which these record values occur are r.v.'s themselves; they are called record times.

A strict definition of record values and record times could be given as follows:

Definition:

The random indices $\{U_n\}_{n=0}^{\infty}$ are inductively defined by

$$U_0(\omega) = 1$$

$$U_n(\omega) = \begin{cases} \min \{k \in \mathbb{N} \mid X_k(\omega) > X_{U_{n-1}(\omega)}(\omega)\}, & \text{if it exists} \\ U_{n-1}(\omega), & \text{otherwise} \end{cases}$$

for $\omega \in \Omega$ and $n \in \mathbb{N}$. $\{U_n\}_{n=0}^{\infty}$ is the sequence of the upper record times and $\{X_{U_n}\}_{n=0}^{\infty}$ the sequence of the upper record values of the sequence $\{X_k\}_{k=1}^{\infty}$. Let $\{L_n\}_{n=0}^{\infty}$ be the sequence of the upper record times of the sequence $\{-X_k\}_{k=1}^{\infty}$. $\{L_n\}_{n=0}^{\infty}$ is called the sequence of lower record times and $\{X_{L_n}\}_{n=0}^{\infty}$ the sequence of lower record values of $\{X_k\}_{k=1}^{\infty}$. □

If the r. v.'s $\{X_k\}_{k=1}^{\infty}$ are independent and identically distributed (i.i.d.) with continuous c.d.f., the sequence of record times is infinite almost surely (a.s.), as can easily be shown by induction. Moreover, there exist infinitely many record values a.s., which finally exceed or go below every fixed value from the interior of the support of the c.d.f. a.s.

As the distribution of the record times then does not depend on the original distribution, they can be used for distribution-free tests in time-series (cf. [2]). In [2], the test-statistics can be expressed in terms of $\max\{k \in \mathbb{Z}^+ | U_k \leq n\}$ and $\max\{k \in \mathbb{Z}^+ | L_k \leq n\}$ for a fixed $n \in \mathbb{N}$ (which are the number of upper and lower records in a series of n observations). Since in stochastic simulation all considered c.d.f.'s are assumed to be known, it is the purpose of this paper to propose a similar test for the independence of random number generators using record values rather than record times. Besides a gain of information using the c.d.f. of $\{X_n\}_{n=1}^{\infty}$, the distributions under consideration are Poisson-distributions while the distributions of the test-statistics in [2] are rather tedious to calculate.

The c.d.f. of the record values can easily be calculated using the following

<u>Lemma:</u>

Let $n \in \mathbb{N}$ and $Y_0, \dots, Y_n$, $Z_1, \dots, Z_n$ be real r.v.'s independent of each other with the c.d.f.'s $F_0, \dots, F_n$ and $G_1, \dots, G_n$ resp. Then

$$P\left(\bigcap_{i=1}^{n} \{Z_i \leq Y_{i-1} \leq Y_i \leq t\}\right) =$$

$$\int_{(-\infty,t]} \int_{(-\infty,t_n]} \cdots \int_{(-\infty,t_1]} \prod_{i=1}^{n} G_i(t_{i-1})\, dF_0(t_0)..dF_n(t_n).$$

<u>Proof:</u>

Let $g: \mathbb{R}^{2n+1} \to \{0,1\}$ be defined by

$$g(t_0, \ldots, t_n, s_1, \ldots, s_n) = \begin{cases} 1, & s_i \leq t_{i-1} \leq t_i \leq t \quad (i=1,\ldots,n) \\ 0, & \text{otherwise} \end{cases}$$

$$= 1_{(-\infty,t]}(t_n) \prod_{j=1}^{n} 1_{(-\infty,t_j]}(t_{j-1}) \prod_{i=1}^{n} 1_{(-\infty,t_{i-1}]}(s_i),$$ where 1_A

denotes the indicator r.v. for any event $A \in \mathfrak{A}$. Then

$$P\left(\bigcap_{i=1}^{n} \{Z_i \leq Y_{i-1} \leq Y_i \leq t\}\right) = \int_{\Omega} 1_{\bigcap_{i=1}^{n}\{Z_i \leq Y_{i-1} \leq Y_i \leq t\}}\, dP =$$

$$\int_{\Omega} g(Y_0, \ldots, Y_n, Z_1, \ldots, Z_n)\, dP = \int_{\mathbb{R}^{2n+1}} g\, dP_{(Y_0,\ldots,Y_n,Z_1,\ldots,Z_n)} =$$

$$\underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{(n+1)\text{-times}} 1_{(-\infty,t]}(t_n) \prod_{j=1}^{n} 1_{(-\infty,t_j]}(t_{j-1}) \cdots$$

$$\underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{n\text{-times}} \prod_{i=1}^{n} 1_{(-\infty,t_{i-1}]}(s_i)\, dG_1(s_1)\ldots dG_n(s_n)\, dF_0(t_0)\ldots dF_n(t_n) =$$

$$\int_{(-\infty,t]} \int_{(-\infty,t_n]} \cdots \int_{(-\infty,t_1]} \prod_{i=1}^{n} G_i(t_{i-1})\, dF_0(t_0) \ldots dF_n(t_n). \quad \square$$

From now, let $\{X_k\}_{k=1}^{\infty}$ be i.i.d. with continuous c.d.f. F. Set $\xi_0 = \inf\{s \in \mathbb{R} \mid F(s) > 0\}$, $\xi_1 = \sup\{s \in \mathbb{R} \mid F(s) < 1\}$. Then we have the following representation:

Lemma:

$$P(X_{U_n} \leq t) = \begin{cases} \int_{-\infty}^{t} \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} \prod_{i=0}^{n-1} \frac{1}{1-F(t_i)} \, dF(t_0) \ldots dF(t_n), & t < \xi_1 \\ 1, & \text{otherwise} \end{cases}$$

for every $n \in \mathbb{N}$.

Proof:

Let $t < \xi_1$ and $k_0 = 1$. Since F is continuous, " $<$ " and " $\leq$ " are interchangeable in the lemma above, and the Monotone Convergence Theorem yields

$$P(X_{U_n} \leq t) = P\Big(\bigcup_{1<k_1<k_2<\ldots<k_n} \bigcap_{m=1}^{n} \{\max_{k_{m-1}<i<k_m} X_i \leq X_{k_{m-1}} < X_{k_m} \leq t\}\Big) =$$

$$\sum_{k_1=k_0+1}^{\infty} \cdots \sum_{k_n=k_{n-1}+1}^{\infty} \int_{-\infty}^{t} \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} \prod_{m=1}^{n} F^{k_m-k_{m-1}-1}(t_{m-1}) \, dF(t_0) \ldots dF(t_n) =$$

$$\sum_{j_1=0}^{\infty} \cdots \sum_{j_n=0}^{\infty} \int_{-\infty}^{t} \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} \prod_{i=0}^{n-1} F^{j_{i+1}}(t_i) \, dF(t_0) \ldots dF(t_n) =$$

$$\int_{-\infty}^{t} \int_{-\infty}^{t_n} \cdots \int_{-\infty}^{t_1} \prod_{i=0}^{n-1} \frac{1}{1-F(t_i)} \, dF(t_0) \ldots dF(t_n) .$$

For $t \geq \xi_1$, $P(X_{U_n} \leq t) = P(X_{U_n} < \infty) = 1$. □

The formula given above can much be simplified by setting

$$R(t) = -\ln(1 - F(t)) = \int_{-\infty}^{t} \frac{1}{1-F(s)} \, dF(s) \quad \text{for} \quad t < \xi_1 :$$

Corollary:

For $t < \xi_1$ and $n \in \mathbb{N}$

$$P(X_{U_n} \leq t) = \int_{-\infty}^{t} \frac{1}{n!} R^n(s)\; dF(s) = \int_{0}^{R(t)} \frac{s^n}{n!} e^{-s}\; ds = 1 - e^{-R(t)} \sum_{k=0}^{n} \frac{R^k(t)}{k!},$$

i.e. $R(X_{U_n})$ follows an Erlang-distribution (c.f. [6], p.69).

This result can readily be obtained by induction. □

Using a standard argument relating the Erlang-distribution with the Poisson-distribution, we are led to the following

Lemma:

The r.v.'s $\{Y_t \mid \xi_0 < t < \xi_1\}$ and $\{Z_t \mid \xi_0 < t < \xi_1\}$, defined by

$Y_t = \min\,\{n \in \mathbb{Z}^+ \mid X_{U_n} > t\}$ and $Z_t = \min\,\{n \in \mathbb{Z}^+ \mid X_{L_n} < t\}$

(with $\min \emptyset = \infty$) are real r.v.'s a.s. following a Poisson-distribution with parameters $R(t)$ and $\overline{R}(t) = -\ln F(t)$ resp.

Proof:

Since F is continuous, Y_t and Z_t are real r.v.'s a.s. (cf. to what has been said after the first definition above). Further,

$$Y_t = n \iff \begin{cases} X_{U_{n-1}} \leq t < X_{U_n}, & n \in \mathbb{N} \\ t < X_1, & n = 0 \end{cases} \quad \text{i.e.}$$

$P(Y_t = 0) = P(X_1 > t) = 1 - F(t) = e^{-R(t)}$ and

$P(Y_t = n) = P(X_{U_{n-1}} \leq t < X_{U_n}) = P(X_{U_n} > t) - P(X_{U_{n-1}} > t) =$

$e^{-R(t)} \frac{R^n(t)}{n!}$ for $n \in \mathbb{N}$ by the corollary and the monotony of record values. The result concerning Z_t follows by transition from $\{X_k\}_{k=1}^{\infty}$ to $\{-X_k\}_{k=1}^{\infty}$, i.e. from F to $1 - F(-\,.)$, and from t to $-t$.

Since $\xi_0 < t < \xi_1$, $0 < R(t), \overline{R}(t) < \infty$, which guarantees that the distributions are not degenerate. □

The r.v.'s Y_t and Z_t can be interpreted as the number of record jumps which are needed to exceed or go below the value t for the first time.

Now let $x_1, \dots, x_n$ ($n \in \mathbb{N}$) be random numbers taken from a random number generator for a c.d.f. F. Choosing ξ_p with $F(\xi_p) = p$ for $0 < p < 1$, a procedure for testing the hypothesis H_0: $x_1, \dots, x_n$ are independent realizations of the random number generator against H_1: H_0 is not true can be described as follows:

Divide $x_1, \dots, x_n$ into finite subsequences with the last element of each subsequence being the first one to exceed or go below ξ_p. For each subsequence, count the number of record jumps (if there is only one element in a subsequence, the number of jumps is zero). Under the assumption of independence these numbers are independent realizations of the r.v.'s Y_{ξ_p} and Z_{ξ_p} resp., which are Poisson-distributed with parameters $\lambda_U = R(\xi_p) = -\ln(1-p)$ and $\lambda_L = \overline{R}(\xi_p) = -\ln p$ resp. The decision between H_0 and H_1 then has to be made depending on the outcome of a goodness-of-fit test such as Pearson's χ^2-test, applied to the numbers of record jumps obtained. Since the mean number of random numbers needed to produce a realization of Y_{ξ_p} and Z_{ξ_p} is Gumbel's return period ([3] p. 21)

$$T(\xi_p) = \frac{1}{1 - F(\xi_p)} = e^{R(\xi_p)} = e^{\lambda_U} = \frac{1}{1-p} \quad \text{and}$$

$$\overline{T}(\xi_p) = e^{\overline{R}(\xi_p)} = e^{\lambda_L} = \frac{1}{p}$$ resp., the mean number of subsequences amounts to $n(1-p)$ and np resp. Obviously, large numbers of subsequences imply small Poisson parameters and vice versa. Since large Poisson parameters yield more information about the original random numbers, p should be chosen accordingly. However, the frequencies with which small record jumps occur should not be less than 5 in order to avoid too much grouping of data concerning the χ^2-test.

Therefore, $n(1-p_U)\, e^{-\lambda_U} = n(1-p_U)^2 \geq 5$

and $n\, p_L\, e^{-\lambda_L} = n\, p_L^2 \geq 5$, i.e. $p_U \leq 1 - \sqrt{\frac{5}{n}}$ and $p_L \geq \sqrt{\frac{5}{n}}$

for upper (p_U) and lower (p_L) records resp.

The testing procedure described above can also be applied when in doubt whether F is the appropriate distribution or when F is completely unknown. This is true, since under the assumption of independence a change of the value of the c.d.f. F at ξ_p yields a corresponding change of Poisson parameter only without leaving the class of Poisson-distributions. In this case λ_U and λ_L are to be estimated by the sample mean after choosing a convenient value for ξ_p. By a well-known theorem of Fisher the χ^2-test then can still be applied having f-1 degrees of freedom instead of f in the usual case.

Example:

The first 5000 random numbers generated by the linear congruential method with the modulus $P = 2^{15}$, the multiplier $\lambda_0 = 899$ and the initial value $r_0 = 3$ (c.f. [4], p. 39/53) give the following results putting $p_U = \frac{15}{16}$ and $p_L = \frac{1}{16}$ (i.e. $\lambda_U = \lambda_L = \ln 16 \sim 2.773$):

	$Y_{\frac{15}{16}}$	$Z_{\frac{1}{16}}$
number of observations (theoretical)	295 (312.5)	319 (312.5)
sample mean	2.759	2.712
frequencies (theoretical)		
0	20 (18.44)	18 (19.94)
1	47 (51.12)	50 (55.28)
2	77 (70.87)	78 (76.63)
3	62 (65.49)	81 (70.82)
4	48 (45.40)	56 (49.09)
5	24 (25.17)	27 (27.22)
6	8 (11.63)	6 (12.58)
7	6 (4.61)	3 (4.98)
8	2 (1.60)	0 (1.73)
9	0 (0.49)	0 (0.53)
10	0 (0.14)	0 (0.15)
11	1 (0.03)	0 (0.04)
χ^2-test statistic T	3.18	9.24
degrees of freedom f	7	7
critical value c at a significance level of 5%	14.07	14.07

In [4], p. 53 the results of several independence tests for the random number generator used above (and others) are given. Among these tests the similar run test with runs above and below the median and the up-and-down run test are of special interest since they deal with the growth behavior of the random numbers just as the record value test does. Further, run tests also use the χ^2-test statistic so that the results are comparable:

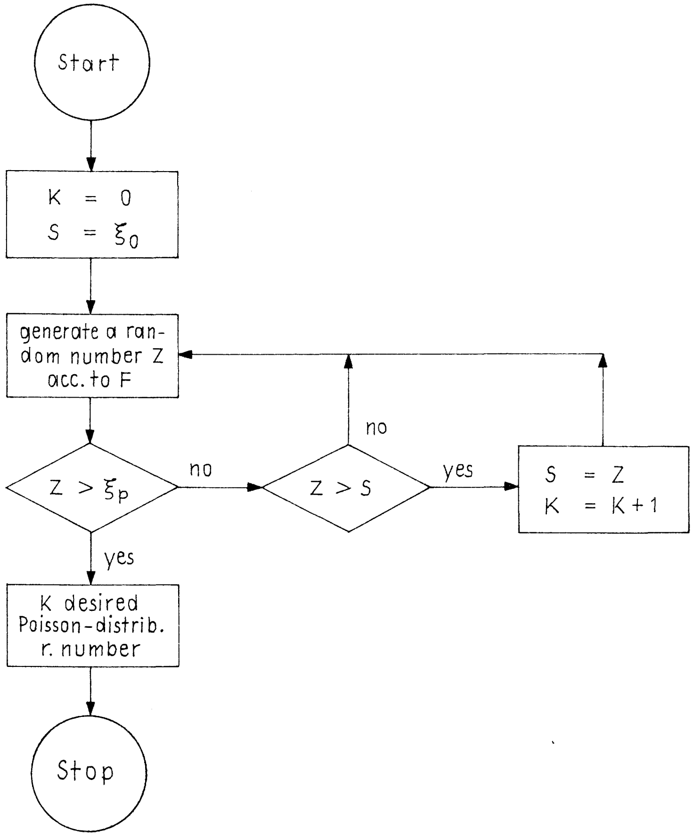
	record value test $p_U = \frac{15}{16}$	$p_L = \frac{1}{16}$	run test with runs above and below the median
T	3.18	9.24	7.09

with $f = 7$ and $c = 14.07$

	record value test $p_U = \frac{1}{2}$	$p_L = \frac{1}{2}$	up-and-down run test
T	2.24	0.23	1.69

with $f = 4$ and $c = 9.49$

The r.v.'s Y_t and Z_t can also be used to generate Poisson-distributed random numbers with parameter $\lambda > 0$ from any continuous c.d.f. F. According to what has been said earlier the algorithm using e.g. upper record values can be described by the following diagram, setting $p = 1 - e^{-\lambda}$:



Since the mean number of random numbers distributed according to F which is needed to produce a Poisson-distributed random number with parameter λ is always e^{λ}, the algorithm might be applicable for small values of λ only; however, for large λ the mean number can be reduced using the fact that the sum of independent Poisson-distributed r.v.'s is again Poisson-distributed. For this purpose, let

$$N = \min \{n \in \mathbb{N} \mid \lambda \leq n(n+1) \ln (1 + \frac{1}{n})\}$$

(which minimizes $f(n) = n\, e^{\lambda/n}$ over $\mathbb{N}$)

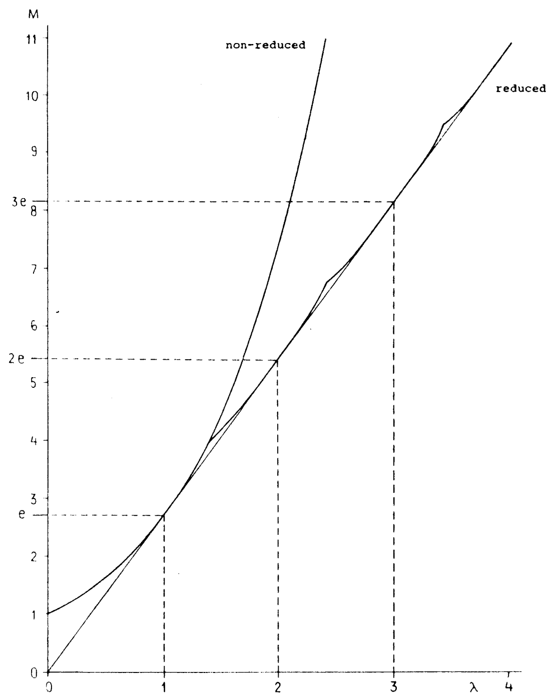
and generate N independent Poisson-distributed random numbers with parameter $\frac{\lambda}{N}$. Summing up these random numbers yields a Poisson-distributed random number with parameter λ. The mean number M of random numbers distributed according to F which is needed then reduces to

$$M = N\, e^{\lambda/N} \leqq N\left(1+\frac{1}{N}\right)^{N+1} \leqq N e + \frac{e}{2} \leqq e\lambda + \frac{3}{2} e$$

since $\lambda > (N-1)N \ln\left(1+\frac{1}{N-1}\right) \geqq (N-1)$ for $N \geqq 2$.

More tedious calculations show that even the inequality

$M \leqq e\lambda + \frac{e}{6(N-1)}$ holds for $N \geqq 2$, i.e. $M - e\lambda \to 0$ if $\lambda \to \infty$.

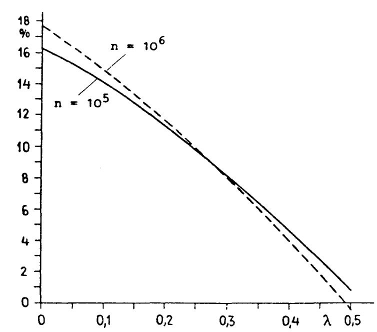


If F is assumed to be the uniform distribution on [0,1] there is another algorithm for generating Poisson-distributed random numbers which is derived from the properties of Poisson processes ([5], p. 172/173). The mean number of uniformly distributed (u.d.) random numbers needed to give a Poisson-distributed random number with mean λ then amounts to $1+\lambda$. A comparison of both methods with respect to CPU-times shows that the algorithm based on records is slightly faster for $\lambda < 0{,}5$ though the mean number of u.d. random numbers needed is $e^{\lambda} > 1+\lambda$. This is mainly due to the simplicity of the algorithm which avoids arithmethic operations on the random numbers needed.

Numerical results were obtained by the aid of a Control Data computer Cyber 175. The relevant steps of the programs for which CPU-times were recorded are given below. Program PROD corresponds to the conventional algorithm, program REC to the algorithm based on record values. The u.d. random numbers are generated by the linear congruential method with modulus 2^{22}, multiplier 648053 and initial value 17. Compilation was done by the FTN compiler with OPT = 2 , XL stands for λ. The percentage of the mean CPU-time saved using REC instead of PROD is shown in the figure below. The dark line and the dashed line correspond to $n = 10^5$ and $n = 10^6$ Poisson-distributed random numbers resp.

```
      PROGRAM PROD                OPT=2
                  .
                  .
                  .
      C=EXP(-XL )
      I=17 $ J=N=0
   10 CONTINUE
      K=0 $ S=1.
   20 CONTINUE
      N=N+1
      I=648053*I
      I=MOD(I,2**22)
      X=I/2.**22
      S=X*S
      IF(S.LT.C)GOTO 30
      K=K+1
      GOTO 20
   30 CONTINUE
      J=J+1
      IF(J.NE.100000)GOTO 10
                  .
                  .
                  .
```

```
      PROGRAM REC                 OPT=2
                  .
                  .
                  .
      C=1.-EXP(-XL )
      I=17 $ J=N=0
   10 CONTINUE
      K=0 $ S=0.
   20 CONTINUE
      I=648053*I
      I=MOD(I,2**22)
      X=I/2.**22
      N=N+1
      IF(X.GT.C)GOTO 30
      IF(X.LE.S)GOTO 20
      S=X
      K=K+1
      GOTO 20
   30 CONTINUE
      J=J+1
      IF(J.NE.100000)GOTO 10
                  .
                  .
                  .
```



References:

[1] CHANDLER, K.N.:
The distribution and frequency of record values. J. Roy. Statist. Soc. Ser. B, 14 (1952), p. 220 - 228

[2] FOSTER, F.G. and A. STUART:
Distribution-free tests in time-series based on the breaking of records. J. Roy. Statist. Soc. Ser. B, 16 (1954), p. 1 - 22

[3] GUMBEL, E.J.:
Statistics of extremes. New York, Columbia University Press, 1958

[4] JÖHNK, M.D.:
Erzeugen und Testen von Zufallszahlen. Berichte aus dem Inst. f. Statistik u. Vers.math. der FU Berlin, 6 (1969), Physica-Verlag Würzburg

[5] MIHRAM, G.A.:
Simulation: Statistical foundations and methodology. New York, Ac. Press, 1972

[6] RESNICK, S.I.:
Limit laws for record values. Stochastic Processes and their Appl. 1 (1973), p. 67 - 82

[7] SHORROCK, R.W.:
On record values and record times. J. Appl. Prob. 9 (1972), p. 316 - 326

Anschrift:
Institut für Statistik und
Wirtschaftsmathematik
RWTH Aachen
5100 Aachen