STATISTICAL ESTIMATION OF THE INDEPENDENCE OF INTERVALS IN DYNAMIC MODELS OF TRAINS OF ACTIONS

SERGEI KOVBASSA

Department of Higher Mathematics,

Saint-Petersburg State University of Economics and Finance

Ul. Sadovaya 21, 191023 Saint-Petersburg, Russia

Tel.(812) 110-5605, Fax (812) 110-5733, e-mail: kovbassa@finec.ru

ABSTRACT

One of basic assumptions of the known dynamic models of the risk theory [1,2] including a sequence of actions during a fixed time interval which is a realization of a random process of occurrence of actions from individual insurance contracts (when several actions can be brought during the contract period, as, for instance, in automobile insurance) is the same distribution and independence of time intervals between the moments the actions are brought. In the paper the statistical method for estimating the degree of correctness of this assumption by using the available empirical data is suggested. To illustrate the method, an example of its use is given.

KEYWORDS

Model of train of actions. The same distribution and independence of random value. Estimation criterion.

1. THEORY

Let us consider how the relationship between trains of actions from two contracts can be estimated. We assume that the investigator has simultaneous records of trains of actions from each individual contract and from two contracts in combination.

Let us designate the first train of actions by X_1 and the second train by X_2. We suppose that X_1 and X_2 are renewal processes, which means that the following assumptions are made:

1) random values $x_1, x_2, \ldots$ (intervals between actions) are independent;

2) integral functions of $x_2, x_3, \ldots$ are the same, i.e., $P(x_k < t) = F(t)$ for $k = 2, 3, 4, \ldots$

Note that it is not necessary that the integral function $F_1(t) = P(x_1 < t)$ be equal to the integral functions of the remaining intervals, i.e., $F_1(t) \neq F(t)$. Function $F(t) = P(x_k < t)$, $k = 2,3, \ldots$ is referred to as the integral function of the renewal process. Let us suppose that the integral function is differentiable and equals zero for negative values of the argument. The derivative of the integral function of the renewal process is called the renewal process density, i.e., if $p(t)=F'(t)$, then $p(t)$ is density. Note that

$$F(t)=\int_0^t p(\Theta)d\Theta,\ t>0$$

Quantity $\mu=\int_{-\infty}^{\infty} tp(t)\mathrm{d}t=\int_0^{\infty} tp(t)\mathrm{d}t$ is the average time between events (actions) of the renewal process.

To find the integral function of the first interval of a stationary ordinary renewal process, we use Statement 1.

Statement 1. Let X be a stationary and ordinary renewal process; $F(t)$ is its integral function; $F_1(t)$ is the integral function of the first interval, i.e., $F_1(t) = P(x_1 < t)$; and μ is the average time between events of the renewal process. Then

$$F_1(t)=\frac{1}{\mu}\int_0^t (1-F(\Theta))\mathrm{d}\Theta$$

Trains X and Y are referred to as independent if random values $x_1, x_2, x_3 \ldots, y_1, y_2, y_3 \ldots$ are independent. Here, $x_1, x_2, \ldots$ are lengths of intervals for train X and $y_1, y_2, \ldots$ are lengths of intervals for train Y.

Let us consider two trains X and Y. We project the moments at which events of trains X and Y happen onto one axis, as shown in Fig.1. The train Z corresponding to the obtained sequence of actions will be referred to as superposition of trains X and Y.

Then Statement 2 is valid.

Statement 2. If X and Y are independent stationary ordinary renewal processes, their superposition is also a stationary renewal process.

At last, Statement 3 is valid.

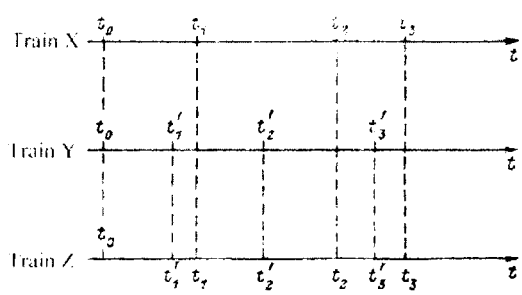

Fig.1. Scheme showing formation of superposition of trains.

t_i, t_i' – moments of occurrence of events (actions).

Statement 3. If X and Y are independent stationary ordinary renewal processes; $F_1^{(X)}$ and $F_1^{(Y)}$ are integral functions of the first time interval for trains X and Y, respectively; and $F_X \oplus F_Y$ is the integral function of the superposition of trains X and Y, we have

$$(F_X \oplus F_Y)(t) = F_1^{(X)}(t) + F_1^{(Y)}(t) - F_1^{(X)}(t) F_1^{(Y)}(t)$$

So Z is assumed to be the superposition of trains X_1 and X_2. Let us designate the integral function of train Z by F_Z. We suppose that trains X_1 and X_2 are independent. Then, as follows from Statement 3, knowing the integral functions F_{X_1} and F_{X_2} of trains X_1 and X_2, it is possible to find the integral function of the superposition of trains under the condition of independence of trains, i.e., $F_{X_1} \oplus F_{X_2}$. Let us label the function thus obtained by $F_{X_1} \oplus F_{X_2}$. It

is evident that if function F_Z is equal to $F_{X_1} \oplus F_{X_2}$, the hypothesis about independence of trains X_1 and X_2 should not be rejected. Therefore, the algorithm of estimation of the interindependence of trains reduces to testing the main hypothesis H.

Hypothesis H. Distributions F_Z and $F_{X_1} \oplus F_{X_2}$ are consistent.

It is natural to apply to this hypothesis the Smirnoff-Kolmogoroff's test. Let us formulate hypothesis H_1: the structure of the train for each contract is described by the following model. If x is the interval between actions in the train, the probability density $f_j(x)$ in this model is given by

$$f_j(x) = \frac{x^{N_j - 1}}{(N_j - 1)!} \cdot \nu_j^{N_j} \exp(-\nu_j x) \qquad (1)$$

Here, $N_j = 1,2, \ldots, \nu_j > 0$, $j = 1,2$. Distribution (1) is the Erlang distribution. There are the following statements related to the Erlang distribution (4 and 5).

Statement 4. Let random value X have the Erlang distribution. Then

$$M(X) = \frac{N}{\nu}, \quad D(X) = \frac{N}{\nu^2}.$$

Here, M(X) is the mathematical expectation, and D(X) is the variance of random value X.

Statement 5. Let X be a stationary ordinary renewal process with the density being the Erlang distribution density with parameters ν and N. For a fixed t we take

$$a_N = F(t) = \int_0^t f_{Erl}(x)\mathrm{d}x$$

$$b_N = \int_0^t (1 - F_{Erl}(x))\mathrm{d}x$$

Here, $N = 1,2, \ldots$ For brevity, the dependence of a_N and b_N on t and ν is not given. Then the integral function of the first time interval is

$$F_1(t) = \frac{\nu}{N} b_N.$$

Parameters a_N and b_N can be found from the recursion formulae

$$a_1 = 1 - e^{-\nu t}, \quad b_1 = \frac{1 - e^{-\nu t}}{\nu},$$

$$a_N = -\frac{(t\nu)^{N-1}}{(N-1)!} e^{-\nu t} + a_{N-1},$$

$$b_N = (1 - a_N)t + \left(\frac{1}{N\nu}\right) a_{N+1}.$$

The correctness of these statements follows from Statement 1. Recursion formulae can be checked by simple calculations. The integral function of the superposition of two independent renewal processes with the Erlang distribution can be found by using Statements 3 and 5.

To test hypothesis H_1, we use the maximum likelihood method and the χ^2 test. It can be assumed that $x_i > 0$ for $i = 1,2, \ldots, n$. Testing of hypothesis H_1 is performed in two stages. The first stage involves finding N^* and ν^* for which the corresponding likelihood function will reach the maximum. The second stage includes testing of hypothesis H_1 with $N = N^*$ и $\nu = \nu^*$ by using the χ^2 test.

Let us describe stages I and II in detail.

Stage I. Finding N^* and ν^*.

It is obvious that, if hypothesis H_1 is satisfied, the likelihood function for a sequence of intervals between actions has the form

$$L(\bar{x}, N, \nu) = \prod_{i=1}^{n} f(x_i) = \frac{\left(\prod_{i=1}^{n} x_i\right)^{N-1}}{((N-1)!)^n} \nu^{nN} \exp\left(-\nu \sum_{i=1}^{n} x_i\right)$$

Here, $\bar{x}$ is the vector with coordinates $x_1, x_2, \ldots x_n$. It is necessary to find the values ν^* and N^* of parameters ν and N at which the likelihood function $L(\bar{x}, N, \nu)$ has a global maximum. Without making special comments, we shall use in further discussion the following designations:

$$r=\prod_{i=1}^{n} x_i;\ s=\sum_{i=1}^{n} x_i;\ g=\frac{s}{\sqrt[n]{r}\,n}.$$

Note that it follows from the classical inequality for arithmetic mean S/n and geometric mean $\sqrt[n]{r}$ that $g \geq 1$; g being equal to 1 only when equality $x_1 = x_2 = \ldots = x_n$ is fulfilled.

Before finding N^* and v^* we note that Statement 6 is valid.

Statement 6. Let $\delta_k = (1 + 1/k)^{k+1}$, where $k = 1, 2, \ldots$

Then

a) if $g > 1$ and $\delta_1 \leq ge$, then $\delta_k \leq ge$ for $k = 1,2,\ldots$;

b) if $g > 1$ and $\delta_1 > ge$, there exists a natural number k_0 such that $\delta_k > ge$ for $1 \leq k < k_0$ and $\delta_j \leq ge$ for $j \geq k_0$ and this number is the only one;

c) if $g = 1$, then $\delta_k > ge$ for $k = 1,2,\ldots$ (here e is the base of the natural logarithm). The correctness of this statement immediately follows from the fact that sequence δ_k monotonically decreases and tends to e for $k \to \infty$.

Then the following statement is valid.

Statement 7.

Case 1. If $g > 1$, the global maximum of the likelihood function $L(\bar{x}, N, v)$ is reached at $v^* = nN^*/S$. In this case

a) if $ge \geq 4$, then $N^* = 1$;

b) if $ge < 4$, then $N^* = k_0$ (see Statement 6).

Case 2. If $g = 1$, the likelihood function is not bounded from above, and therefore a global maximum does not exist.

Proof. Let us fix N. Then the likelihood function $L(\bar{x}, N, v)$ will be the function of one argument. Let us find the local maximum of this function ($v > 0$). Note that with the designations used the likelihood function acquires the form

$$L(\bar{x}, N, v) = \frac{r^{N-1}}{((N-1)!)^n} v^{nN} \exp(-vs)$$

Let us calculate the derivative $L_v'(\bar{x}, N, v)$. It is evident that

$$L'_v(\bar{x},N,v)=\frac{r^{N-1}}{((N-1)!)^n}v^{nN-1}\exp(-vs)(nN-sv)$$

This means that at fixed N the local maximum of the $L(\bar{x},N,v)$ function is reached at point $v^* = nN/s$. Since the local maximum occurs at the only point v^*, the global maximum of the likelihood function is also reached at point v^* in case N is fixed.

It is obvious that if the global maximum of the $L(\bar{x},N,v)$ function with respect to variables N and v is reached, it takes place at $v = v^*$.

Thus the problem reduces to finding the maximum (over all natural N) of the following sequence

$$\alpha_N = L(\bar{x},N,v^*)=\frac{r^{N-1}(nN)^{nN}}{((N-1)!)^n s^{nN}}\exp(-nN)$$

With the designations used, we have

$$\alpha_k=\frac{r^{-1}}{((k-1)!)^n}\left(\frac{k}{ge}\right)^{kn}$$

It is easy to see that

$$\frac{\alpha_{k+1}}{\alpha_k}=\left(\frac{\delta_k}{ge}\right)^n$$

Let us recall that

$$\delta_k=\left(1+\frac{1}{k}\right)^{k+1}$$

Case 1 is a simple consequence of Statement 7. Let us consider case 2. Let $g = 1$. Then

$$\frac{\alpha_{k+1}}{\alpha_k}=\left(\frac{\delta_k}{e}\right)^n$$

Since $\delta_k > e$ for $k = 1,2,\ldots$, the sequence α_k monotonically increases. By using the known Stirling formula, it can easily be shown that

$$\alpha_k = \frac{rk^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{\Theta n}{12k}\right), (0<\Theta<1)$$

Hence, $\alpha_k \to \infty$ for $k \to \infty$. Statement 7 is proved.

Corollaries to Statement 7.

1. If $g \geq 4/e \approx 1.4715$, then $N^* = 1$.

2. If $1.2416 \approx 27/8e \leq g < 4/e \cong 1.4715$, then $N^* = 2$.

Comments: a) if $g = 1$, then $x_1 = x_2 = \ldots = x_n$. However, the probability of event ($X_1 = X_2 = \ldots = X_n$) is zero since the regular train corresponding to the case $x_1 = x_2 = \ldots = x_n$ is not stationary. Therefore, g is always more than 1; b) in practice the inequality $g \geq 1.2416$ is typically fulfilled, whence $N^* = 2$ or $N^* = 1$.

Having estimated N^* и $v^* = N^* n / \sum_{i=1}^{n} x_i$, we pass to the second stage.

Stage II. Testing of hypothesis H_1 by using the χ^2 test.

Table 1

Sequence of intervals between actions for different contracts

(sample size $n = 50$; the table should be read along the lines)

Trains	Lengths of Intervals (relative units)
X	61.1, 67.9, 48.4, 47.1, 1.2, 20.0, 17.1, 27.7, 15.4, 73.4, 29.2, 30.3, 19.6, 5.3, 89.1, 21.3, 18.9, 37.2, 30.1, 3.2, 7.7, 35.1, 42.2, 21.4, 21.2, 10.2, 59.2, 64.3, 71.5, 51.2, 14.3, 9.8, 24.3, 48.2, 41.2, 36.5, 13.2, 3.1, 54.1, 18.9, 11.1, 39.2, 4.9, 30.2, 16.1, 35.6, 34.4, 43.1, 53.4, 25.9.
Y	49.3, 62.3, 45.5, 38.2, 21.3, 36.6, 40.2, 8.9, 4.5, 7.8, 28.1, 32.7, 83.5, 42.2, 78.8, 23.8, 53.3, 70.0, 66.6, 47.1, 18.3, 1.2, 19.9, 39.4, 68.7, 7.6, 17.2, 46.8, 55.1, 98.2, 3.3, 16.8, 13.4, 18.3, 26.5, 27.3, 26.6, 30.1, 9.9, 5.0, 21.3, 25.0, 2.3, 34.5, 11.3, 19.9, 27.9, 17.6, 15.4.
Z	20.7, 19.9, 2.2, 39.9, 2.4, 8.2, 23.4, 14.9, 31.2, 3.7, 15.1, 11.1, 1.1, 28.8, 32.4, 16.6, 21.2, 1.1, 54.7, 54.7, 72.1, 66.7, 89.3, 13.1, 13.2, 46.4, 23.9, 9.9, 7.3, 15.5, 69.5, 5.2, 47.8, 3.3, 3.3, 32.6, 44.4, 9.7, 19.2, 21.2, 20.5, 13.4, 6.5, 13.0, 46.4, 38.1, 28.9, 18.8, 8.4, 4.4

This hypothesis is tested in a usual fashion.

II. EXAMPLE

Let us give the example demonstrating the suggested procedure. Trains of actions for the first and second contracts and also for both contracts in combination are shown in Table 1 by samples X, Y, Z, respectively. It is necessary to estimate the degree of interrelation between trains X and Y.

For train X we have $\sum_{i=1}^{50} x_i = 1605$; $\left(\prod_{i=1}^{50} x_i\right)^{1/50} = 23.8975$; $g = 1.3432$. Therefore, according to the comment to Statement 3, we obtain $N^* = 2$, $\nu^* = 0.0629$.

Similarly, for train Y we have $\sum_{i=1}^{50} y_i = 1602$, $\left(\prod_{i=1}^{50} y_i\right)^{1/50} = 22.3596$, $g = 1.4329$ и $N^* = 2$; $\nu^* = 0.0624$.

Then we apply the χ^2 test to the hypothesis H_1 that X and Y are distributed in accordance with the Erlang law with parameters $N^* = 2$ and $\nu^* = 0.06$. Calculations are given in Table 2. We choose the significance level $\alpha = 0.05$. Since $\chi^2_{X,Y} < \chi^2_{0.05}$ ($\chi^2_{0.05} = 14.1$ is chosen from the table of distribution of χ^2 for seven degrees of freedom and significance level $\alpha = 0.05$, and $\chi^2_X = 8.80$ and $\chi^2_Y = 8.21$ are found from Table 2), hypothesis H_1 is not rejected.

Let us find the integral function for trains X and Y by using results of Statements 3 and 5. As a result, we have

$$(F_x \oplus F_y)(x) = 1 - 0.5(1 + 0.06x)(2 + 0.06x)\exp(-0.12x).$$

Table 2

Results of applying the χ^2 test to hypothesis H_1

Boundaries of intervals (rel.units)	Frequencies m_i		Estimate of np_i^*	$m_i - np_i^*$		$\frac{(m_i - np_i^*)^2}{np_i^*}$	
	X	Y		X	Y	X	Y
0 - 10	7	9	6.000	1.000	3.000	0.0275	0.2500
10 - 20	10	10	11.000	-1.000	-1.000	0.0081	0.0081
20 - 30	9	9	10.000	-1.000	-1.000	0.0100	0.0100
30 - 40	8	7	7.500	0.500	-0.500	0.0044	0.0044
40 - 50	6	6	5.500	0.500	0.500	0.0081	0.0081
50 - 60	4	2	3.500	0.500	-1.500	0.0200	0.1836
60 - 70	3	3	2.670	0.330	0.330	0.0140	0.0140
70 - 80	2	2	1.500	0.500	0.500	0.1100	0.1100
80 - 90	1	1	0.235	0.760	0.760	7.6000	7.6000
90 - 100	0	1	0.865	-0.865	0.135	1.1000	0.0243
Σ	50	50	—	—	—	—	—

Note: $\chi_X^2 = 8.80$, $\chi_Y^2 = 8.21$.

At last, we apply the Smirnoff-Kolmogoroff's test to the hypothesis H about consistency of sample Z with distribution $F_X \oplus F_Y$. Calculations are given in Table 3.

Table 3

Results of applications of Smirnoff-Kolmogoroff's test to hypothesis

H

Boundaries of intervals (rel. units)	Frequencies	Cumulative frequencies	Integral functions		$\lvert F_n(x) - (F_x \oplus F_y) \rvert$
			empirical $F_n(x)$	theoretical $F_x \oplus F_y$	
0 - 10	15	15	0.3000	0.3800	0.0800
10 - 20	12	27	0.5400	0.6800	0.1400
20 - 30	8	35	0.7000	0.9100	0.2100
30 - 40	5	40	0.8000	0.9330	0.1330
40 - 50	4	44	0.8800	0.9750	0.0950
50 - 60	2	46	0.9200	0.9904	0.0704
60 - 70	2	48	0.9600	0.9964	0.0364
70 - 80	1	49	0.9800	0.0087	0.0187
80 - 90	1	50	1.0000	0.9991	0.0109
90 - 100	0	50	1.0000	0.9999	0.0001
Σ	50	—	—	—	—

Since

$$D_{50}^{(0)} = \max \lvert F_n(x) - (F_x \oplus F_y)(x) \rvert = 0.210$$

then $\lambda_0 = 50^{1/2} \cdot 0.210$. Let us choose the significance level $\alpha = 0.05$. Since $1 - K(1.487) = 0.0236 < 0.05$, this means that an improbable event has occurred. For this reason the hypothesis is rejected, i.e., trains X and Y are interrelated. This conclusion is made with a reliability of 0.95.

REFERENCES

1. C.D. Daykin, T. Pentikainen and M. Pesonen. Practical Risk Theory for Actuaries. Chapman & Hall, 1994.

2. G.I.Falin. Mathematical Analysis of Risks in Insurance. Russian Law Publishing Co., Moscow, 1994.