ACTA MATHEMATICA VIETNAMICA

Vol. 13 No 1 (1988)

ON THE STATISTICAL ANALYSIS OF A RANDOM NUMBER OF OBSERVATIONS

NGUYEN BAC VAN

1. INTRODUCTION

There exist many practical situations in which the observations are made at random times, and the number of these observations is random. For example, rainfalls are observed at their random happening moments; and in a fixed time interval, the number of observations, or that of rainfalls, is random. The following model will be appropriate for such situations.

Let $(T_n, n \geqslant 1)$ be a stricly increasing unbounded sequence of nonnegative random variables, possibly taking infinite values.

$(T_n, n \geqslant 1)$ is called a simple point process (non-explosive) on the extended real semiline. It generates a counting process $(N(t), t \geqslant 0)$ by

$$N(t) = n \text{ if } T_n \leqslant t < T_{n+1}$$

Then $\lim\limits_{t \to +\infty} N(t) = +\infty$ iff all terms T_n are finite.

Let $(X_n, n \geqslant 1)$ be a sequence of real-valued (for simplicity) random variables. The double sequence $(T_n, X_n, n \geqslant 1)$ is callled a marked point process. In a fixed observation interval $[0, t]$, we get a sequence of mark observations $(X_n, 1 \leqslant n \leqslant N(t))$. (1)

On the basis of the data (1), we shall evaluate the common probability distribution of the variables X_n, once they are identically distributed, and estimate moments of this distribution.

Notations and abbreviations: In the sequel, $(\Omega, \mathcal{F}, P)$ denotes the basic probability space, ω element of Ω, I_A the indicator of a set A, E or E_P the expectation with respect to the probability measure P, $E(Y|Z)$ the conditional expectation of Y given z, a. s. stands for « almost surely », P for « in probability ». Let the realisation of a relation R depends on the element ω; let A be the set of all elements ω at which R is realized, A can belong to $\mathfrak{F}$ or not. Then R is said to hold a. s. if A includes some almost sure event B, i. e. $A \overset{a.s.}{\supset} B$, $B \in \mathfrak{F}$ and $PB = 1$.

2. PROPOSITIONS

LEMMA 1. *Consider two real-valued random functions* $(Y_r, r \in Q)$, $(N(t), t \in S)$, *the second Q-valued ; Q and S are upper boundless sets in* R^1. *Suppose that*

$$Y_r \xrightarrow{a.s.} Y \text{ as } r \to +\infty \ (r \in Q),$$

where Y *is some measurable function on* Ω

$$N(t) \xrightarrow{a.s.} +\infty \text{ as } t \to +\infty \ (t \in S).$$

Then

$$Y_{N(t)} \xrightarrow{a.s.} Y \text{ as } t \to +\infty \ (t \in S).$$

When Q and S are the set$_s$ of positive integers, this lemma becomes Theorem 1 in [1].

Proof. A, B, C denote subsets of Ω defined as follows

$A = (Y_{N(t)}$ does not tend to Y as $t \to +\infty)$,

$B = (N(t)$ does not tend to $+\infty$ as $t \to +\infty)$, $(t \in S)$,

$C = A(\Omega - B)$,

$D = (Y_r$ does not tend to Y as $r \to +\infty)$, $(r \in Q)$.

Then $A \subset B + C$. (a)

Take an arbitrary element $\omega \in C$. Because $\omega \in A$, for $\varepsilon > 0$, there exists in S a sequence $s_k \to +\infty$ such that for every k

$$|Y_{N(s_k)}(\omega) - Y(\omega)| \geqslant \varepsilon.$$

$N(s_k) \to +\infty$ because $\omega \in \Omega - B$.

Hence $\omega \in D$. i.e. $C \subset D$. (b)

The assertion of Lemma 1 is derived from (a), (b) and from the fact that B and D are included in some null events.

Before going further, let us recall some concepts used in [2].

For $Z = \varphi(X_1, X_2, \ldots)$, the translate by $k - 1$ $(k \geqslant 1)$ is defined as follows:

$$Z_k = \varphi(X_k\ X_{k+1}, \ldots).$$

A Borel function fo $(X_n, n \geqslant 1)$ is called invariant, if it coincides with all its translates. If all such invariant functions degenerate into. a.s. constants, the sequence $(X_n, n \geqslant 1)$ is called indecomposable.

PROPOSITION 1. *Let Z be a real-valued Borel fu.c.ion of the family $(X_n, n \geqslant 1)$, with $E\,|Z| < \infty$, and Z_k the translate by $k-1$ of Z. If*

$$\left.\begin{array}{l}\text{the sequence } (X_n,\ n \geqslant 1) \text{ is stationary} \\ \text{and indecomposable,} \\ \text{and } N(t) \xrightarrow{P} +\infty \text{ as } t \to +\infty,\end{array}\right\} \quad (2)$$

then

$$N^{-1}(t) \sum_{k=1}^{N(t)} Z_k \xrightarrow{a.s.} EZ \text{ as } t \to +\infty \quad (3)$$

Proof. $N(t) = \sum_{k=1}^{\infty} I(T_k \leqslant t)$ is $\mathcal{P}$ - measurable.

Because $N(t) \xrightarrow{P} +\infty$ as $t \to +\infty$, there exists an increasing sequence $t_k \to +\infty$, such that $N(t_k) \xrightarrow{a.s} +\infty$ as $k \to \infty$.
Because $N(t)$ is nondecreasing in t, it follows that

$$N(t) \xrightarrow{a.s} +\infty \quad \text{as } t \to \infty \quad (4)$$

Let us set

$$\begin{cases} Y_r = r^{-1} \sum_{k=1}^{r} Z_k \text{ for } r \geqslant 1, \\ Y_0 = C = \text{an arbitrary constant.} \end{cases}$$

By the ergodic theorem ([2], § 30.4),

$$Y_r \xrightarrow{a.s.} EZ \text{ as } r \to +\infty \quad (5)$$

Then (3) follows from (4), (5) and Lemma 1. Q.E.D.

Setting sucessively
$Z = f(X_1)$, a real-valued Borel function of X_1,
$Z = I_{(X_1 \in S)}$, where S is a Borel set in R^1,
we obtain

COROLLARY 1. *If* (2) *is satisfied, then* $t \to a\,s + \infty$

a) $$N^{-1}(t) \sum_{k=1}^{N(t)} f(X_k) \xrightarrow{a.s.} Ef(X_1)$$

provided $E\, f(X_1) < +\infty$,

b) $$\frac{N_S(t)}{N(t)} \xrightarrow{a.\,s.} P(X_1 \in S)$$

where

$$N_S(t) = \sum_{k=1}^{N(t)} I_{(X_k \in S)} \quad (6)$$

PROPOSITION 2. *Let* $F(x) = P(X_1 < x)$,

$$F_{N(t)}(x) = N^{-1}(t) \sum_{k=1}^{N(t)} I_{(X_k < x)}$$

Then, if (2) is satisfied,

$$\sup_{-\infty<x<+\infty} | F_{N(t)} - F(x) | \xrightarrow{a.s.} 0$$

as $t \to +\infty$.

Proof. Noticing the fact that, if $G(x)$, $F(x)$ are two probability distribution functions (left-continuous), we have for every $r = 1, 2, \ldots$:

$$\sup_{-\infty<x<+\infty} | G(x) - F(x) | \leqslant$$

$$\leqslant \frac{1}{r} + \max_{k=1,\ldots,r} \Big[| G(x_{rk}) - F(x_{rk}) | , | G(x_{rk} + 0) - F(x_{rk} + 0) | \Big] \quad (7)$$

where x_{rk} $(k = 1 \ldots, r)$ are defined by

$$x_{rk} = \inf \left(x : F(x) \leqslant \frac{k}{r} \leqslant F(x + 0) \right).$$

Consider the sets A_{rk}, A'_{rk} $(k = 1, \ldots, r;\ r = 1, 2, \ldots)$ in Ω, defined by

$$A_{rk} = \{ F_{N(t)}(x_{rk}) \to F(x_{rk}) \text{ as } t \to +\infty \},$$

$$A'_{rk} = \{ F_{N(t)}(x_{rk} + 0) \to F(x_{rk} + 0) \text{ as } t \to +\infty \},$$

and the countable intersection

$$\bigcap_{r=1}^{\infty} \bigcap_{k=1}^{r} A_{rk} A'_{rk}.$$

Replacing $G(x)$ in (7) by $F_{N(t)}(x)$, we see that (8) is contained in the set

$$A = \{ \sup_{-\infty<x<+\infty} | F_{N(t)}(x) - F(x) | \to 0 \text{ as } t \to +\infty \}.$$

Let $S = (-\infty, x_{rk})$ or $(-\infty, x_{rk}]$ in Corollary (1b) then we see also that A_{rk}, A_{rk} include almost sure events, so does (8), and hence A. Q.E.D.

We now pass to a parameter estimation. The underlying family of probability measures will be some class $\mathcal{P}$ of probability distributions of the family $(N(t), X_n, t \geqslant 0, n \geqslant 1)$. On the basis of the data (1), any estimator g of a vector parameter $m = m(P)$, $(P \in \mathcal{P})$, taking values in a space R^d, is *a* R^d — valued function defined on Ω by means of $X_1(\omega), \ldots, X_{N(t)}(\omega)$, i.e,

$$g(X_1, \ldots, X_{N(t)}) = \sum_{n=0}^{+\infty} g_n(X_1, \ldots, X_n) I_{Q_n} \quad (9)$$

where $Q_n = \{ \omega : N(t) = n \}$. On the set $Q_n (n \geqslant 0)$, the function g reduces to a function $g_n(X_1,\dots,X_n)$. The $g'_n s(n \geq 1)$ are supposed to be R^d — valued Borel functions of $X_1,\dots,X_n$, while g_0 is a constant.

In practice, once the observations in a fixed time interval $[0, t]$ have been collected, one always draws statistical inference in some definite situation $N(t) = n$. So, we introduce the.

DEFINITION 1. A sequence $\{g_n(X_1,\dots,X_n), n \geq 1\}$ is called a conditionally unbiased *estimator* for the vector parameter m, if for every $n \geq 1$ such that $PQ_n \geq 0$,

$$E_p\{g_n(X_1,\dots,X_n) \mid Q\} = m$$

for all $P \in \mathcal{P}$.

PROPOSITION 3. *Let* $(N(t) \to +\infty$ *as* $t \to +\infty$. *Then any conditionally unbiasedestimator defines an asymtotically (as* $t \to \infty$*) unbiased estimate of* m*, according to* (9)*, with an arbitrary constant* g_0 .

Proof. From (9), by the σ-additivity of the indefinite integral

$$\int_{\Omega} g(X_1,\dots,X_{N(t)})\, dP = \sum_{n=0}^{\infty} \int_{Q_n} g_n(X_1,\dots,X_n)\, dP$$

provided the left side integral exists.

If $PQ_n > 0$, we get

$$\int_{Q_n} g_n(X_1,\dots,X_n)\, dP = PQ_n \cdot E\{g_n(X_1,\dots,X_n) | Q_n\}$$

Hence, from (10) and by Definition 1, for all $P \in \mathcal{P}$,

$$E_P\, g(X_1,\dots,X_{N(t)}) = g_0\, PQ_0 + m \sum_{n=1}^{\infty} PQ_n$$

$$= m + (g_0 - m)\, PQ_0 .$$

It follows that, for all $P \in \mathcal{P}$,

EXAMPLE 1. Let $(X_n, n \geq 1)$ be stationary. Suppose that X_k and $N(t)$ $I_{(N(t) \geq k)}$ are independent for every $k \geq 1$. Let $f(X_1)$ be any real-valued Borel function of X_1, with $E_p\, |f(X_1)| < \infty$. Then,

$$\overline{f_t} = \begin{cases} N^{-1}(t) \sum_{k=1}^{N(t)} f(X_k) & \text{when } N(t) > 0, \\ C = \text{const.} & \text{when } N(t) = 0 \end{cases}$$

is a conditionally unbiased estimator for the parameter $m = E_p\, f(X_1)$.

Indeed, when $PQ_n > 0$, $n \geqslant 1$,

$$E_P(\bar{f}_t \mid Q_n) = E_P(n^{-1} \sum_{k=1}^{n} f(X_k) \mid Q_n)$$
$$= E_P f(X_1) = m,$$

because, by Assumption, $f(X_k)$ and $Q_n = N(t) = n$ are independent for $n \geqslant k \geqslant 1$, hence

$$E_p(f(x_k) \mid Q_n) = E_P(f(x_k)) = E_P f(x_1).$$

Note that, without any independence between X_k and $N(t)$, by Corollary 1a, f_t is a strongly consistent estimate for $m = E_P f(X_1)$.

Finally, in view of an impostant practical application, we give the following proposition, slightly generalizing the Wald's lemma in [3] (4. 4), with a simpler proof.

PROPOSITION 4. *Let $(X_n, n \geqslant 1)$ be stationary, X_k and $I_{(N(t) \geqslant k)}$ be independent for each $k \geqslant 1$. Let $f(X_k)$ be any realvalued Borel function of X_k. We make the convention that*

$$\sum_{k \leqslant N(t)} f(X_k) = 0 \text{ when } N(t) = 0$$

If $f \geqslant 0$, then

$$E\{\sum_{k \leqslant N(t)} f(x_k)\} = E N(t). E f(x_1) \tag{11}$$

If f has an arbitrary sign, but $E \mid f(X_1) \mid < \infty$, $EN(t) < \infty$, then (11) still holds.

Proof. $\sum_{k \leqslant N(t)} f(X_k) = \sum_{k=1}^{+\infty} f(X_k) I_{(N(t) \geqslant k)}$

If $f \geqslant 0$, then by the theorem of the monotone convergence, we have:

$$E\{\sum_{k \leqslant N(t)} f(X_k)\} = \sum_{k=1}^{\infty} E\{f(X_k) I_{(N(t) \geqslant k)}\} =$$

$$= \sum_{k=1}^{\infty} Ef(X_k). EI_{(N(t) \geqslant k)} = Ef(X_1). \sum_{k=1}^{\infty} P(N(t) \geqslant k) = Ef(X_1). E N(t) \tag{12}$$

If f has an arbitrary sign, the above equalities are valid for $|f|$ instead of f. Hence

$$E \sum_{k=1}^{\infty} |f(X_k)| I_{(N(t) \geqslant k)} = E \mid f(X_1) \mid . E N(t) < +\infty$$

Then, the equalities (12) hold by the Lebesgue dominated convergence theorem.

Q.E.D.

In particular,

$$E N_S(t) = P(X_1 \in S). EN(t) \tag{13}$$

where $N_S(t)$ is given by (6).

3. APPLICATION

In building sciences, if X_k is the intensity (defined in a suitable manner) of the k th rainfall since the moment zero, the critical level of rainfall is defined as a constant a_t such that the event $(X_k \geqslant a_t)$ occurs in the average one time during the time interval $[0, t]$.

By setting $S = [a_t, +\infty]$ in (13), we get the important formula

$$P(X_1 \geqslant a_t) = \frac{1}{EN(t)}$$

In the problem of evacuation of rain-water for towns, one needs the expression $a_t = a(t)$. If for $x > 0$, $F(x) = P(X_1 < x)$ is strictly increasing in x, then (14) gives

$$a_t = \overset{-1}{F}\left[1 - \frac{1}{EN(t)}\right]$$

By Proposition 2, using the empirical distribution function $F_{N(t)}(x)$, we can seek a suitable form for F(x), If this form contains some unknown moments, we can estimate them as in Example 1. In [4], details of this statistical analysis are exposed on the basis of rainfall's intensity observations, collected by the Central Meteorological Station in Hanoi, and an approximate numerical formula for $a_t = a(t)$ is obtained.

REFERENCES

[1] W. Richter, *Übertragung von Grenzaussagen für Folgen von zufälligen Grössen auf Folgen mit zufälligen Indizes*, Teoriya Veroyatnost. Primen., X, 1 (1965), 82 – 94.

[2] M. Loeve, *Probability theory*, 3rd ed., D. Van Nostrand, 1963.

[3] S. Zacks, *The theory of statistical inference*, Wiley Sons, 1971.

[4] Nguyen Bac Van, *On the method of statistical analysis of rainfall's intensity observations (in Vietnamese)*, Report given at the Research Station of Meteorological Sciences in Hanoi, 8-1975 (unpublished)

Received January 30, 1987

UNIVERSITY OF HOCHIMINH CITY