Natural Science Report, Ochanomizu University, Vol. 46, No. 2, 1995
お茶の水女子大学 自然科学報告 第46巻 第2号

CONDITIONALLY TRIMMED SUMS FOR INDEPENDENT RANDOM VARIABLES

Yuji KASAHARA[1]

Department of Information Sciences, Ochanomizu University

(Received October 10, 1995)

Abstract

This paper studies conditionally trimmed sums for triangular arrays of independent random variables and generalize the Hahn-Kuelbs result for i.i.d. cases.

1990 Mathematics Subject Classification: Primary 60F17, 60F05; Secondary 62G30.

1. Introduction

Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with common nondegenerate distribution function $F(x)$, and let $\{X_j^{(n)}\}_{j=1}^n$ $(n \geq 1)$ be the order statistics based on the sample $\{X_1, \dots, X_n\}$ in the descending order in absolute value; i.e., $\{X_j^{(n)}\}_{j=1}^n$ is a rearrangement of $\{X_j\}_{j=1}^n$ so that

$$|X_1^{(n)}| \geq |X_2^{(n)}| \geq \cdots \geq |X_n^{(n)}|.$$

Let p_n $(n = 1, 2, \dots)$ be a sequence of positive, nondecreasing integers such that $1 \leq p_n \leq n$ and define

$$\begin{aligned} S_n^{p_n} &= \sum_{j=p_n+1}^{n} X_j^{(n)} \\ &= \sum_{j=1}^{n} X_j - \sum_{j=1}^{p_n} X_j^{(n)}, \qquad n = 1, 2, \dots \end{aligned}$$

$S_n^{p_n}$, which is often referred to as the *trimmed sum*, is the n-th partial sum with the p_n largest samples trimmed and its asymptotic distribution as $n \to \infty$ have been studied by many authors. (e.g. Stigler [4], Mori [3], etc. See Hahn-Kuelbs [2] for more references.)

A common interest of these authors is to get asymptotic normality theorems by deleting extreme samples even in the case where the tail probability of X_j is large (and hence the CLT no longer holds for the usual partial sums), and it is one of the important philosophy that it is desirable to retain as much data as possible. From this point of view, Hahn-Kuelbs ([2]) introduced the notion of *conditionally trimmed sums*; let a_n $(n \geq 1)$ be positive numbers and let

$$S_n^{p_n}(a_n) = \sum_{j=1}^{n} X_j - \sum_{j=1}^{p_n} X_j^{(n)} I(|X_j^{(n)}| > a_n).$$

Hence $S_n^{p_n}(a_n)$ denotes the n-th partial sum with the p_n largest samples trimmed provided that they exceed a_n in magnitude. Hahn-Kuelbs ([2]) proved that for arbitrary nondegenerate distribution function $F(x)$,

[1]Supported in part by Grant-in-Aid for Scientific Research(No. 07640284), Ministry of Education, Science and Culture.

we can find p_n, a_n, b_n and c_n such that $p_n/n \to 0, a_n \to \infty$ as $n \to \infty$ and that $(1/c_n)\{S_n^{p_n}(a_n) - b_n\}$ converges in law to a Gaussian random variable. In fact, the existence of such sequences p_n, a_n, b_n and c_n is easy itself (see Remark in Section 3). So the significance of their result is offering an explicit procedure to find p_n, a_n, b_n and c_n. Now the aim of the present paper is to consider a similar problem for sums of triangular arrays of independent random variables. The author admits that all necessary ideas are found in [2], but we believe that their idea will be clearer under our formulation.

We give the main theorem in Section 2 and the i.i.d. case will be discussed in Section 3.

2. Main Theorem

For every $n = 1, 2, \ldots$, let $\{\xi_{n,1}, \xi_{n,2}, \ldots, \xi_{n,k_n}\}$ be a sequence of independent random variables ($k_n \to \infty$) and let $\{\tilde{\xi}_{n,1}, \tilde{\xi}_{n,2}, \ldots, \tilde{\xi}_{n,k_n}\}$ be a rearrangement of the samples $\{\xi_{n,1}, \xi_{n,2}, \ldots, \xi_{n,k_n}\}$ so that

$$|\tilde{\xi}_{n,1}| \geq |\tilde{\xi}_{n,2}| \geq \cdots \geq |\tilde{\xi}_{n,k_n}|.$$

For every p $(1 \leq p \leq k_n)$ and $a(> 0)$, we define the conditionally trimmed sum $S_n^p(a)$ as follows:

$$S_n^p(a) = \sum_{j=1}^{n} \xi_{n,j} - \sum_{j=1}^{p} \tilde{\xi}_{n,j} I(|\tilde{\xi}_{n,j}| > a).$$

Thus $S_n^p(a)$ denotes the sum of $\{\xi_{n,j}\}$ with the p largest samples deleted provided that they exceed the prescribed level $a\,(> 0)$ in magnitude.

Theorem 1. *Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive numbers and let*

$$v_n = \sum_j \mathrm{Var}\big(\xi_{n,j} I(|\xi_{n,j}| \leq a_n)\big).$$

If

$$\text{(A.1)} \qquad \sum_j P\big(\varepsilon\sqrt{v_n} < |\xi_{n,j}| \leq a_n\big) \longrightarrow 0 \ (n \to \infty), \quad \text{for every } \varepsilon > 0,$$

then, for any $p_n \in \mathbb{Z}_+$ $(n \geq 1)$ satisfying the condition

$$\text{(A.2)} \qquad \frac{1}{p_n} \sum_j P\big(|\xi_{n,j}| > a_n\big) \longrightarrow 0,$$

it holds

$$Z_n := \frac{1}{\sqrt{v_n}}\left[S_n^{p_n}(a_n) - m_n\right] \xrightarrow{\mathcal{L}} Z$$

where Z is an $N(0,1)$–random variable and

$$m_n = \sum_j E\big[\xi_{n,j} I(|\xi_{n,j}| \leq a_n)\big].$$

("$\xrightarrow{\mathcal{L}}$" denotes the convergence in law.) A sufficient condition for (A.1) is

$$\text{(A.3)} \qquad a_n = o\big(\sqrt{v_n}\big) \qquad \text{as } n \to \infty.$$

Proof. Let

$$S_n^* = \sum_{j=1}^{k_n} \xi_{n,j} I\big(|\xi_{n,j}| \leq a_n\big).$$

We shall first see that

$$Z_n^* := \frac{1}{\sqrt{v_n}}\{S_n^* - m_n\} \xrightarrow{\mathcal{L}} Z.$$

This can easily be seen because $S_1^*, S_2^*, \ldots$ are usual sums of independent random variables. Indeed, it can be rewritten as follows;

$$Z_n^* = \sum_j (\zeta_{n,j} - E[\zeta_{n,j}])$$

where

$$\zeta_{n,j} = \frac{1}{\sqrt{v_n}}\xi_{n,j} I(|\xi_{n,j}| \le a_n).$$

First, it holds

$$\begin{aligned}\sum_j P(|\zeta_{n,j}| > \varepsilon) &= \sum_j P\Big(\frac{1}{\sqrt{v_n}}|\xi_{n,j}| I(|\xi_{n,j}| \le a_n) > \varepsilon\Big) \\ &= \sum_j P\big(|\xi_{n,j}| > \varepsilon\sqrt{v_n} \quad \text{and} \quad |\xi_{n,j}| \le a_n\big)\end{aligned}$$

which vanishes as $n \to \infty$ by (A.1). Furthermore, by the definition of v_n, we also see that

$$\sum_j \mathrm{Var}(\zeta_{n,j}) = 1.$$

Therefore, by the usual CLT, we obtain

$$Z_n^* \xrightarrow{\mathcal{L}} Z. \tag{2.1}$$

We next show that $P(|Z_n - Z_n^*| > \varepsilon) \to 0$, for every $\varepsilon > 0$:

$$\begin{aligned}P(|\tilde{\xi}_{n,p_n}| > a_n) &= P(\#\{j : |\xi_{n,j}|\} \ge p_n) = P\Big[\frac{1}{p_n}\sum_j I[|\xi_{n,j}| > a_n] \ge 1\Big] \\ &\le E\Big[\frac{1}{p_n}\sum_j I[|\xi_{n,j}| > a_n]\Big] = \frac{1}{p_n}\sum_j P(|\xi_{n,j}| > a_n)\end{aligned}$$

which converges to 0 by (A.2). Thus we obtain

$$P(|\tilde{\xi}_{n,p_n}| > a_n) \to 0.$$

Since $S_n^{k_n}(a_n) = S_n^*$ on the event $\{|\tilde{\xi}_{n,p_n}| \le a_n\}$ it holds that

$$P[Z_n = Z_n^*] \ge P[|\tilde{\xi}_{n,p_n}| \le a_n] \longrightarrow 1. \tag{2.2}$$

Combining (2.1) and (2.2) we obtain the assertion of the theorem. □

3. The Case of IID Random Variables

In this section we shall study the case where the triangular array $\{\xi_{n,j}\}$ comes from an i.i.d. sequence: Let $X_1, X_2, \ldots$ be a nondegenerate i.i.d. random variables with the common distribution function F as in Introduction and let $k_n = n$, $\xi_{n,j} = X_j$. Then, as a special case of Theorem 1 in the previous section we have the following theorem, which is is a modification of Hahn-Kuelbs ([2]).

Theorem 2. *Let a_n be a sequence tending to infinity and one of the following two conditions is satisfied.*

(B.1a)
$$\frac{n}{a_n^2}\int_{|x|\le a_n} x^2\,dF(x) \longrightarrow \infty \quad (n\to\infty)$$

(B.1b)
$$n\int_{|x|>\varepsilon\sqrt{n}} dF(x) \longrightarrow 0 \quad (n\to\infty), \qquad \text{for every } \varepsilon>0.$$

Then, for any p_n $(1\le p_n\le n)$ such that

(B.2)
$$\frac{n}{p_n}\int_{|x|>a_n} dF(x) \longrightarrow 0 \quad (n\to\infty),$$

it holds that

(2.1)
$$\frac{1}{b_n}\{S_n^{p_n}(a_n)-c_n\} \xrightarrow{\mathcal{L}} Z \ (n\to\infty)$$

where

$$b_n=\sqrt{n\int_{|x|\le a_n} x^2\,dF(x)},$$
$$c_n=n\int_{|x|\le a_n} x\,dF(x)$$

and Z is an $N(0,1)$–random variable.

Proof. Since v_n in Theorem 1 is equal to $n\int_{|x|<a_n} x^2\,dF(x)$, (B.1a) is equivalent to (A.3). If (B.1b) is satisfied instead, then it is easy to see that $nP[|X_1|>\varepsilon\sqrt{v_n}]\to 0$, which implies (A.1). Furthermore, (A.2) can be rewritten as (B.2). Thus we have the assertion. □

Remarks. (i) We can always find a_n $(\uparrow\infty)$ satisfying (B.1a) unless $P[X_1=0]=1$. Also we can choose p_n so that $p_n=o(n)$ in (B.2).

(ii) A sufficient condition for (B.1b) is $E[X_1^2]<\infty$ since, in general,

$$nP[|X/\sqrt{n}|>\varepsilon]\le \varepsilon^{-2}E[X^2 : |X|>\varepsilon\sqrt{n}].$$

(iii) As we mentioned in Introduction, if we are interested only in the existence of $\{p_n,\ a_n,\ b_n,\ c_n\}$ satisfying (2.1), the proof is easier: For every fixed $a>0$, choose $\alpha>0$ small enough so that $|X_n^{[\alpha n]}|>a$ for all sufficiently large n (a.s.), which is possible due to the law of large numbers. Then, for all sufficiently large n, $S_n^{[\alpha n]}(a)$ is equal to $Y_1+\cdots+Y_n$ where $Y_k=X_kI[|X_k|\le a]$. Therefore, under a suitable linear normalization, $S_n^{[\alpha n]}(a)$ converges in law to an $N(0,1)$-random variables. Now by a standard argument, choosing α_n tending slowly enough, we can find $a_n(\uparrow\infty)$, b_n and c_n satisfying (2.1).

References

[1] S. Csörgő, E. Haeusler and D. M. Mason, The asymptotic distribution of trimmed sums, Ann. Probab. **16**(1988) 672–699.

[2] M. G. Hahn and J. Kuelbs, Universal asymptotic normality for conditionally trimmed sums, Statis. Probab. Letters **7**(1989) 9-15.

[3] T. Mori, On the limit distribution of lightly trimmed sums, Math. Proc. Cambridge Philos. Soc. **96**(1984) 507-516.

[4] S. M. Stigler, The asymptotic distribution of the trimmed mean. Ann. Statist. **1**(1973) 472–477.