Ann. Inst. Statist. Math.
Vol. 54, No. 1, 114–124 (2002)

EXPANSIONS FOR THE DISTRIBUTIONS OF SOME NORMALIZED SUMMATIONS OF RANDOM NUMBERS OF I.I.D. RANDOM VARIABLES

NAN WANG[1] AND WEI LIU[2]

[1]*Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, U.K.,*
e-mail: nw@socsci.soton.ac.uk
[2]*Department of Statistics, University of Central Florida, Orlando, FL, U.S.A. and*
Department of Mathematics, University of Southampton, Southampton, SO17 1BJ, U.K.,
email: wl@maths.soton.ac.uk

(Received September 8, 1999; revised June 12, 2000)

Abstract. The central limit theorem for a normalized summation of random number of i.i.d. random variables is well known. In this paper we improve the central limit theorem by providing a two-term expansion for the distribution when the random number is the first time that a simple random walk exceeds a given level. Some numerical evidences are provided to show that this expansion is more accurate than the simple normality approximation for a specific problem considered.

Key words and phrases: Central limit theorem, expansion of a tail probability, martingale, renewal theory, sequential analysis, stopping time, Wald's lemma.

1. Introduction

The central limit theorem for a summation of random number of i.i.d. random variables appears in 1950's (see e.g. Rényi (1957)). It plays an important role in sequential analysis. However asymptotic normality is often not accurate enough. In many problems, more accurate and reliable approximations are desirable. Although there are well established results such as Berry-Esseen theorem and Edgeworth expansions when the number of random variables in a summation is non-random, little is known when the number of random variables in a summation is random (see discussion later).

In Rényi's (1957) central limit theorem, the random number is an integer valued random variable, τ_a, such that τ_a/a converges in probability to a positive constant as $a \to \infty$. Unfortunately, under this general setting, not much progress can be made to improve the central limit theory (see e.g. Landers and Rogge (1976) and Ghosh *et al.* (1997), Chapter 2). So we consider τ_a that is a stopping time having the form

$$\tau_a = \inf\{n \geq 1 : S_n > a\}, \qquad a \geq 1$$

where $S_n = X_1 + \cdots + X_n$ and $\{X_n, n \geq 1\}$ are i.i.d. random variables with $E(X_1) = 1$. Note that τ_a is the first time that the simple random walk $\{S_n\}$ exceeds level a. It has applications in areas such as queuing theory or insurance risk theory (see Gut (1988)) and in sequential data analysis (see Woodroofe (1982)).

Let $\{Y_n, n \geq 1\}$ be another sequence of i.i.d. random variables such that $E(Y_1) = 0$ and $(X_1, Y_1), (X_2, Y_2), \ldots$ are independent random vectors. In this paper, some two-term

expansions for the distributions of

$$A_1 := \frac{Y_1 + \cdots + Y_{\tau_a}}{\sqrt{a}} \quad \text{and} \quad A_2 := \frac{Y_1 + \cdots + Y_{\tau_a}}{\sqrt{\tau_a}}$$

are established as $a \to \infty$ in the manner: $a = j\delta$ and $j \to \infty$ where δ is an arbitrarily fixed positive number. The major tool used is a result of Mykland (1993) which provides a two-term expansion for a martingale. Note that this expansion is slightly different from Edgeworth expansion in that its remainder approaches zero in the sense of Mykland (1993) but not in the usual sense for an Edgeworth expansion.

Edgeworth-type expansions for the distributions of summations of random number of i.i.d. random variables are available in the literature for only three special situations. The first one assumes the independence between τ_a and $Y_1 + \cdots + Y_{\tau_a}$ (see e.g. Bose and Boukai (1996)), for which the problem is reduced essentially to the Edgeworth expansion for a sum of fixed-number of i.i.d. random variables by conditioning on τ_a. The second, considered by Woodroofe and Keener (1987) and extended to the vector situation by Lai and Wang (1994), assumes a special relationship between X_i and Y_i. The third special case, discussed by Takahashi (1987), assumes not only a special relationship between X_i and Y_i but also the normality of the distribution of Y_i. The techniques employed in these special cases are not applicable to our setting where no special relationship between X_i and Y_i is assumed.

The layout of the paper is as follows. In Section 2, a martingale sequence induced by the stopping time τ_a is introduced and some of its properties are discussed. In Section 3, a result of Mykland (1993) and the properties established in Section 2 are used to provide two-term expansions for the distributions of A_1 and A_2. In Section 4, the accuracy of the expansions is assessed and compared with the simple normality approximation in a particular problem.

2. A martingale induced by τ_a

The expansions in Theorem 3 of this paper are established when $a \to \infty$ in the manner: $a = j\delta$ and $j \to \infty$ where $\delta > 0$ is arbitrarily fixed. This allows the expansions to be used for any large a, either integer or non-integer. For notational simplicity, we assume in the proofs of Sections 2 and 3 that $a \in N^+$, the set of all positive integers. (Otherwise, the τ_i below should be defined by $\inf\{n \geq 1 : S_n > i\delta\}$.) Assume that $0 < E(X_1^2) < \infty$ and $E(Y_1^2) = 1$. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$, the σ-algebra generated by $\{X_1, \ldots, X_n, Y_1, \ldots, Y_n\}$. Define $\tau_0 = 0$ and

$$\tau_i = \inf\{n \geq 1 : S_n > i\}, \qquad i \in N^+.$$

Then $\{\tau_i\}$ is a sequence of nondecreasing stopping times with respect to $\{\mathcal{F}_n,\ n \geq 1\}$. By Stein's lemma (see e.g. Gut (1988) Theorem 3.1), $P\{\tau_i < \infty\} = 1$ for all i.

LEMMA 1. *If $E|X|^r < \infty$ for $r \geq 1$, then $E(\tau_i - \tau_{i-1})^r$, $i \geq 1$ are uniformly bounded.*

PROOF. Write $\tau_i - \tau_{i-1}$ as

$$\tau_i - \tau_{i-1} = \begin{cases} 0 & \text{if } \ i - (X_1 + \cdots + X_{\tau_{i-1}}) < 0, \ \text{otherwise} \\ \inf\{n \geq 1 : X_{\tau_{i-1}+1} + \cdots + X_{\tau_{i-1}+n} > i - (X_1 + \cdots + X_{\tau_{i-1}})\}. \end{cases}$$

By noting that $X_1 + \cdots + X_{\tau_{i-1}} > i - 1$ for $i \geq 2$ and the fact that *for any $\mathcal{F}_n$-stopping time τ that satisfies $P\{\tau < \infty\} = 1$, $\{X_{\tau+n}, n \geq 1\}$ and $\{Y_{\tau+n}, n \geq 1\}$ are i.i.d. (independent of $\mathcal{F}_\tau$, the σ-algebra prior to the stopping time τ) and each has the same distribution as that of X_1 and Y_1 respectively* (see e.g. Chow and Teicher (1978)), one gets $E(\tau_1)^r \geq E(\tau_i - \tau_{i-1})^r$ for $r \geq 1$ and $i \geq 2$. The lemma then follows since $E(\tau_1)^r$ is bounded by Stein's lemma (see e.g. Gut (1988)).

The statement in italic above is used several times in the sequel. Denote for $i \geq 1$

$$U_i = Y_{\tau_{i-1}+1} + \cdots + Y_{\tau_i} \quad \text{if} \quad \tau_i > \tau_{i-1} \text{ and } 0 \text{ otherwise}$$

and $\mathcal{F}_{\tau_i}$ the σ-algebra prior to the stopping time τ_i.

LEMMA 2. *Suppose $E(Y_1)^{2r} < \infty$ for some $r \geq 1$. Then there is a constant C_0 such that $E(U_i)^{2r} \leq C_0$ for all $i \geq 1$.*

PROOF. By Theorem 5.1 of Gut (1988) and the remark immediately after this proof, $E(U_i)^{2r} \leq CE(\tau_i - \tau_{i-1})^r$, where C is a constant not dependent on $\{\tau_i, i \geq 1\}$. The result then follows from Lemma 1.

Remark. Let τ and s be two stopping times with respect to $\{\mathcal{F}_n, n \geq 1\}$. Then $\tau \vee s - \tau \wedge s$ is a stopping time with respect to $\{\mathcal{B}_n := \mathcal{F}_{\tau \wedge s+n}, n \geq 1\}$. So Theorem 5.1 of Gut (1988) is still valid for $(Y_{\tau \wedge s+1} + \cdots + Y_{\tau \vee s})^{2r}$.

LEMMA 3. *Let $T_i = U_1 + \cdots + U_i$. Then $\{T_i, \mathcal{F}_{\tau_i}, i \geq 1\}$ is a zero-mean martingale.*

PROOF. The lemma follows directly from Doob's optional stopping theorem (see e.g. Woodroofe (1982)) if we can show that

(i) $E|Y_1 + \cdots + Y_{\tau_i}| < \infty$ for each $i \geq 1$;

(ii) $\liminf_{n\to\infty} E|(Y_1 + \cdots + Y_{\tau_i})I_{\{\tau_i > n\}}| = 0$ for each $i \geq 1$.

Note that $\{\tau_i/i, i \geq 1\}$ is uniformly integrable by Scheffé's theorem (see e.g. Rao (1984)) since $\tau_i/i \to 1$ a.s. and $E(\tau_i/i) \to 1$ as $i \to \infty$ from the renewal theory. So, by Lemma 5 of Chow and Yu (1981), $\{(Y_1 + \cdots + Y_{\tau_i})^2/i, i \geq 1\}$ is uniformly integrable and, as a result, $E|Y_1 + \cdots + Y_{\tau_i}| < \infty$ for each fixed i. Part (ii) follows from Holder inequality

$$E|(Y_1 + \cdots + Y_{\tau_i})I_{\{\tau_i > n\}}| \leq \left(i \cdot P\{\tau_i > n\} \cdot E \left| \frac{Y_1 + \cdots + Y_{\tau_i}}{\sqrt{i}} \right|^2 \right)^{1/2}$$

and then Lemma 5 of Chow and Yu (1981) and Stein's lemma.

LEMMA 4. *Let $V_i = U_1^2 + \cdots + U_i^2, i \geq 1$ Then $\{V_i - \tau_i, \mathcal{F}_{\tau_i}, i \geq 1\}$ is a zero-mean martingale.*

PROOF. Given $\tau_{i-1} = m \in N^+$ and $S_{\tau_{i-1}} = R \in (-\infty, \infty)$, define

$$d = \begin{cases} \inf\{n \geq 1 : X_{m+1} + \cdots + X_{m+n} > 1 - R\}, & \text{if} \quad 1 - R \geq 0 \\ 0, & \text{if} \quad 1 - R < 0. \end{cases}$$

Note that, conditioning on $\mathcal{F}_{\tau_{i-1}}$, $d = \tau_i - \tau_{i-1}$. Then by the second order Wald's lemma $E[(X_{m+1}+\cdots+X_{m+d})^2 - d] = 0$ and so $E\{[U_i^2 - (\tau_i - \tau_{i-1})] \mid \mathcal{F}_{\tau_{i-1}}\} = 0$. This completes the proof.

LEMMA 5. *Let $R_i = X_1 + \cdots + X_{\tau_i} - i$ be the overshoot at the time τ_i. Then $\{R_i, i \geq 1\}$ is a time homogeneous Markov chain. Further, if the distribution of X_1 is nonarithmetic, then Markov chain $\{R_i, i \geq 1\}$ has an invariant distribution with density*

$$\lim_{i\to\infty} P\{R_i \in (x, x+dx)\}/dx = H(dx)/dx = \frac{1}{E(S_\tau)} P\{S_\tau > x\}, \qquad x > 0$$

where $\tau = \inf\{n \geq 1 : S_n > 0\}$ and $S_\tau = X_1 + \cdots + X_\tau$.

The proof of the lemma is omitted since a similar proof can be found in Asmussen ((1987), Chapter 4, Proposition 1.5). It is also noteworthy that

$$E[f(U_i, d_i) \mid \mathcal{F}_{\tau_{i-1}}] = E[f(U_i, d_i) \mid R_{i-1}] \tag{2.1}$$

for any Borel measurable function f such that $f(U_i, d_i)$ is integrable, since the distribution of U_i depends on $\mathcal{F}_{\tau_{i-1}}$ only through d_i.

THEOREM 1. *If $E(Y_1^4) < \infty$ then*

$$\sum_{i=1}^{n} U_i^2/n \xrightarrow{P} 1 \qquad as\ n \to \infty.$$

Furthermore, if the distribution of X_1 is nonarithmetic, then

$$\frac{\sum_{i=1}^{n}[U_i^2 - (\tau_i - \tau_{i-1})]^2}{n} \xrightarrow{P} M \qquad as \quad n \to \infty$$

where M is a constant.

PROOF. First we consider

$$\sum_{i=1}^{n}[U_i^2 - (\tau_i - \tau_{i-1})]/n.$$

By Lemma 4, $U_i^2 - (\tau_i - \tau_{i-1})$, $i \geq 1$ are martingale increments and so $E[U_i^2 - (\tau_i - \tau_{i-1})][U_j^2 - (\tau_j - \tau_{j-1})] = 0$ for $i \neq j$. It follows therefore from Chebyshev's inequality that

$$\sum_{i=1}^{n}[U_i^2 - (\tau_i - \tau_{i-1})]/n \xrightarrow{P} 0 \qquad \text{as} \quad n \to \infty.$$

This and the fact that $\tau_n/n \to 1$ a.s. as $n \to \infty$ from the renewal theory imply

$$\sum_{i=1}^{n} U_i^2/n \xrightarrow{P} 1 \qquad \text{as} \quad n \to \infty.$$

To prove the second limit, we consider

$$\sum_{i=1}^{n}\{(U_i^2-(\tau_i-\tau_{i-1}))^2-E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid\mathcal{F}_{\tau_{i-1}}]\}/n.$$

By noting that $(U_i^2-(\tau_i-\tau_{i-1}))^2-E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid\mathcal{F}_{\tau_{i-1}}]$, $i\geq 1$ are martingale increments, a similar argument as above gives

$$\sum_{i=1}^{n}[U_i^2-(\tau_i-\tau_{i-1})]^2/n-\sum_{i=1}^{n}E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid\mathcal{F}_{\tau_{i-1}}]/n\xrightarrow{P}0. \tag{2.2}$$

Also note $E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid\mathcal{F}_{\tau_{i-1}}]=E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid R_{i-1}]$ by (2.1) and $E[(\tau_i-\tau_{i-1})^2\mid R_{i-1}]$, as a function of R_{i-1}, is bounded (see Lemma 1). So there exists a bounded Borel measurable function $f(x)$ such that $E[(U_i^2-(\tau_i-\tau_{i-1}))^2\mid\mathcal{F}_{\tau_{i-1}}]=f(R_{i-1})$. But the ergodic property of a Markov chain (see e.g. Theorem 3.6, Chapter 4 of Revuz (1975)) gives

$$\sum_{i=1}^{n}f(R_{i-1})/n\to\int f(x)H(dx)\quad\text{a.s.}\quad\text{as } n\to\infty$$

which, together with (2.2), implies the required limit.

3. Two-term expansions

Mykland (1993) developed a two-term expansion for a double array of martingale increments. This result is used in this section to provide two-term expansions for the distributions of A_1 and A_2. We first give Mykland's result below.

Suppose $\{X_{nk},\mathcal{F}_{nk},1\leq k\leq k_n,n\geq 1\}$ be a double array of martingale increments such that $E(X_{n(k+1)}\mid\mathcal{F}_{nk})=0$. Denote

$$F_n(x)=P\{S_n/\sqrt{\sigma_n}\leq x\}$$

$$\underbrace{[S,\ldots,S]_{nk}}_{l\text{times}}=\sum_{i=1}^{k}X_{ni}^l$$

$$\underbrace{\langle S,\ldots,S\rangle_{nk}}_{l\text{times}}=\sum_{i=1}^{k}E(X_{ni}^l\mid\mathcal{F}_{n(i-1)})$$

where $S_n=\sum_{k=1}^{k_n}X_{nk}$, and $\sigma_n>0$ is a normalizing factor which is allowed to be random. Let c_n be a sequence of constants such that $c_n^{1/2}=O_P(S_n)$. The following four conditions are required:

(C1) $E[S,S,S,S]_{nk_n}=O(c_n^2/n)$;

(C2) there is a constant b such that, for $(S,S)_{nk_n}$ being either $[S,S]_{nk_n}$ or $\langle S,S\rangle_{nk_n}$, $\sqrt{n}(\frac{(S,S)_{nk_n}}{c_n}-b^2)$ is uniformly integrable;

(C3) there are constants b_* and $\delta>0$ such that $\sup_n E(\sqrt{n}|\frac{\sigma_n}{c_n}-b_*^2|)^{1+\delta}<\infty$;

(C4) there are Borel measurable functions ψ_0, ψ_p and ψ_* so that, if

$$\left\{b^{-1}\frac{S_n}{\sqrt{c_n}},\sqrt{n}\left(\frac{[S,S]_{nk_n}}{c_n}-b^2\right),\sqrt{n}\left(\frac{\langle S,S\rangle_{nk_n}}{c_n}-b^2\right),\sqrt{n}\left(\frac{\sigma_n}{c_n}-b_*^2\right)\right\}$$
$$\xrightarrow{p}(Z,\xi_0,\xi_p,\xi_*)\quad\text{as } n\to\infty$$

then

$$E(\xi_0 \mid Z) = b^2\psi_0(Z) \quad \text{a.s.}$$
$$E(\xi_p \mid Z) = b^2\psi_p(Z) \quad \text{a.s.}$$
$$E(\xi_* \mid Z) = b_*^2\psi_*(Z) \quad \text{a.s.}$$

THEOREM 2. (Mykland (1993)) *Suppose* $\{X_{nk}, \mathcal{F}_{nk}, 1 \le k \le k_n, n \ge 1\}$ *be a double array of martingale increments such that* $E(X_{n(k+1)} \mid \mathcal{F}_{nk}) = 0$ *and the conditions* (C1–C4) *are satisfied. Then, for a twice differentiable function* $g(x)$, *the following expansion is valid*

$$\int_{-\infty}^{+\infty} g(x)dF_n(x) = \int_{-\infty}^{+\infty} g(x)d\Phi(\beta^{-1}x) + \frac{1}{\sqrt{n}}J(g) + o\left(\frac{1}{\sqrt{n}}\right) \tag{3.1}$$

as $n \to \infty$, *where* $\beta = bb_*^{-1}$ *and*

$$J(g) = \frac{1}{2}E\left[\beta^2\left(\frac{1}{3}\psi_0 + \frac{2}{3}\psi_p\right)g''(\beta Z) - \psi_*(Z)\beta Z g'(\beta Z)\right].$$

The convergence in (3.1) *is uniform on* $\mathcal{C}$, *a class of functions* $\{g\}$ *with* g, g' *and* g'' *uniformly bounded, and with* $\{g'', g \in \mathcal{C}\}$ *equicontinuous a.e. Lebesgue.*

Mykland (1993) also gave the following alternative form for expansion (3.1):

$$F_n(x) = \Phi(\beta^{-1}x) + \frac{1}{2\sqrt{n}}\lambda(\beta^{-1}x)\phi(\beta^{-1}x) + o_2\left(\frac{1}{\sqrt{n}}\right) \tag{3.2}$$

where Φ and ϕ are the cdf and pdf of $N(0,1)$ respectively,

$$\lambda(x) = \left(\frac{1}{3}\psi_0'(x) + \frac{2}{3}\psi_p'(x)\right) - \left(\frac{1}{3}\psi_0(x) + \frac{2}{3}\psi_p(x)\right)x + \psi_*(x)x$$

and $o_2(n^{-1/2})$ means $n^{1/2}o_2(n^{-1/2})$ is small in the following sense of Mykland ((1992), Remark 2.3). *A class of finite variation functions* $\{G_a, a \in [1,\infty)\}$ *is written as* $G_a = o_2(a^{-1/2})$ *if*

$$\sup_{g\in C}\left|\sqrt{a}\int_{-\infty}^{\infty} g(x)dG_a(x)\right| \to 0 \quad as \quad a \to \infty$$

where C *is any class of twice continuously differentiable functions* $\{g\}$ *such that* $|g(x)|$, $|g'(x)|$ *and* $|g''(x)|$ *are uniformly bounded for all* $x \in (-\infty,\infty)$ *and that* $\{g'' : g \in C\}$ *is equicontinuous.* It is not clear which, between $o_2(n^{-1/2})$ and the usual $o(n^{-1/2})$, is stronger.

From Section 2, the numerator of A_1 and A_2 is a sum of martingale increments. So we only need to verify the conditions (C1–C4) in order to establish the expansions for the distributions of A_1 and A_2 by using Theorem 2. In the following, we verify (C1–C4) only for A_2, since the conditions for A_1 can be checked in a similar way. The assumptions that the distribution of X_1 is nonarithmetic, $E(X_1)^r < \infty$ for some $r > 2$ and $E(Y_1)^4 < \infty$ are required throughout. Notationwise, we replace n in Mykland's result by a, and denote $X_i' = X_i - 1$ and $S_i' = X_1' + \cdots + X_i'$.

(C1) Set $k_a = c_a = a$. (if $a = j\delta$ then set $k_j = j$ and $c_j = j\delta$.) Then, from Lemma 2, $E(U_1^4 + \cdots + U_a^4) \le aC_0$ and so $E(U_1^4 + \cdots + U_a^4) = O(a) = O(c_a^2/a)$ as required.

(C2) Set $b = 1$ then

$$\sqrt{a}\left(\frac{[S,S]_{ak_a}}{c_a} - b^2\right) = \frac{\sum_{i=1}^a U_i^2 - a}{\sqrt{a}} = \frac{\sum_{i=1}^a [U_i^2 - (\tau_i - \tau_{i-1})]}{\sqrt{a}} + \frac{(\tau_a - a)}{\sqrt{a}}.$$

Note that $(\tau_a - a)/\sqrt{a}$ is uniformly integrable since

$$\frac{\tau_a - a}{\sqrt{a}} = \frac{R_a - S'_{\tau_a}}{\sqrt{a}}, \tag{3.3}$$

R_a^r is uniform integrable in a by Theorem 2.4 of Woodroofe (1982), and $(S'_{\tau_a}/\sqrt{a})^r$ is uniformly integrable by Lemma 5 of Chow and Yu (1981), where $r > 1$ is a constant. Note also that $\{\sum_{i=1}^a [U_i^2 - (\tau_i - \tau_{i-1})]/\sqrt{a}\}$ is uniformly integrable, since $E\{\sum_{i=1}^a [U_i^2 - (\tau_i - \tau_{i-1})]/\sqrt{a}\}^2$ is uniformly bounded by Lemmas 1, 2 and 4. This verifies the uniform integrability of $\sqrt{a}([S,S]_{ak_a}/c_a - b^2)$.

To show the uniform integrability of $\sqrt{a}(\langle S,S\rangle_{ak_a}/c_a - b^2)$, we consider

$$\begin{aligned} Q_1(a) :&= \sqrt{a}\left(\frac{[S,S]_{ak_a}}{c_a} - b^2\right) - \sqrt{a}\left(\frac{\langle S,S\rangle_{ak_a}}{c_a} - b^2\right) \\ &= \sum_{i=1}^a \frac{U_i^2 - E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})}{\sqrt{a}}. \end{aligned}$$

From this, it is clear that $Q_1(a)$ is uniformly integrable, since $E[Q_1(a)^2]$ is uniformly bounded by Lemma 2 and the fact that $\{U_i^2 - E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})\}$ are martingale increments. The uniform integrability of $\sqrt{a}(\langle S,S\rangle_{ak_a}/c_a - b^2)$ now follows from the uniform integrability of $Q_1(a)$ and $\sqrt{a}([S,S]_{ak_a}/c_a - b^2)$. This completes the verification of (C2).

(C3) Set $\sigma_a = \tau_a$ and $b_* = 1$. Then (C3) is clear since $[(\tau_a - a)/\sqrt{a}]^r$ is uniformly integrable for some $r > 1$ from the proof of (C2).

(C4) Consider the linear combination of the four components

$$\begin{aligned} Q_2(a) :&= l_1\frac{\sum_{i=1}^a U_i}{\sqrt{a}} + l_2\frac{\sum_{i=1}^a U_i^2 - a}{\sqrt{a}} + l_3\frac{\sum_{i=1}^a E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}}) - a}{\sqrt{a}} + l_4\frac{\tau_a - a}{\sqrt{a}} \\ &= l_1\frac{\sum_{i=1}^a U_i}{\sqrt{a}} + l_2\left(\frac{\sum_{i=1}^a [U_i^2 - (\tau_i - \tau_{i-1})]}{\sqrt{a}} + \frac{\tau_a - a}{\sqrt{a}}\right) \\ &\quad + l_3\left(\frac{\sum_{i=1}^a U_i^2 - a}{\sqrt{a}} - \frac{\sum_{i=1}^a [U_i^2 - E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})]}{\sqrt{a}}\right) + l_4\frac{\tau_a - a}{\sqrt{a}} \\ &= l_1\frac{\sum_{i=1}^a U_i}{\sqrt{a}} + (l_2 + l_3)\frac{\sum_{i=1}^a [U_i^2 - (\tau_i - \tau_{i-1})]}{\sqrt{a}} \\ &\quad - l_3\left(\frac{\sum_{i=1}^a [U_i^2 - E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})]}{\sqrt{a}}\right) + (l_2 + l_3 + l_4)\frac{R_a - S'_{\tau_a}}{\sqrt{a}}, \end{aligned}$$

where l_1, l_2, l_3 and l_4 are constants, and the last equality uses (3.3). Denote for $i \ge 1$

$$W_i = X'_{\tau_{i-1}+1} + \cdots + X'_{\tau_i} \quad \text{if} \quad \tau_i > \tau_{i-1} \text{ and } 0 \text{ otherwise, and}$$
$$D_i = l_1 U_i + (l_2 + l_3)[U_i^2 - (\tau_i - \tau_{i-1})] - l_3[U_i^2 - E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})] - (l_2 + l_3 + l_4)W_i.$$

Then

$$Q_2(a) = \frac{\sum_{i=1}^{a} D_i}{\sqrt{a}} + \frac{(l_2 + l_3 + l_4)R_a}{\sqrt{a}}. \tag{3.4}$$

By noting that $\{\sum_{i=1}^{a} D_i, \mathcal{F}_a, a \geq 1\}$ is a zero-mean martingale and that $E[D_i^2 \mid \mathcal{F}_{\tau_{i-1}}] = E[D_i^2 \mid R_{i-1}]$, similar arguments as in the second part of the proof of Theorem 1 yield

$$\frac{\sum_{i=1}^{a} D_i^2}{a} \xrightarrow{P} C \quad \text{as} \quad a \to \infty \tag{3.5}$$

where $C = C(l_1, l_2, l_3, l_4)$ is a constant and depends on $\boldsymbol{l}^T = (l_1, l_2, l_3, l_4)$ as a quadratic form, $\boldsymbol{l}^T(\sigma_{ij})\boldsymbol{l}$ say where (σ_{ij}) is a symmetric 4×4 matrix. So $\sum_{i=1}^{a} D_i/\sqrt{a}$ converges in distribution to $N(0, C)$ by the martingale central limit theorem (see e.g. Theorem 3.2 of Hall and Heyde (1980)). It is apparent that $R_a/\sqrt{a}$ goes to zero in probability, since R_a has a limit distribution. Therefore, from (3.4), $Q_2(a)$ converges in distribution to $N(0, \boldsymbol{l}^T(\sigma_{ij})\boldsymbol{l})$ for any $\boldsymbol{l}$, and so

$$\left\{ \frac{\sum_{i=1}^{a} U_i}{\sqrt{a}}, \frac{\sum_{i=1}^{a} U_i^2 - a}{\sqrt{a}}, \frac{\sum_{i=1}^{a} E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}}) - a}{\sqrt{a}}, \frac{\tau_a - a}{\sqrt{a}} \right\} \xrightarrow{D} (Z, \xi_0, \xi_p, \xi_*) \quad \text{as} \quad a \to \infty \tag{3.6}$$

where (Z, ξ_0, ξ_p, ξ_*) is multivariate normal with mean vector zero and covariance matrix (σ_{ij}).

Recall that the conditional expectation $E[X \mid Y] = Y\sigma_{XY}/\sigma_Y^2$ if (X, Y) is bivariate normal with $E(X) = E(Y) = 0$, $\mathrm{Var}(Y) = \sigma_Y^2$ and $\mathrm{Cov}(X, Y) = \sigma_{XY}$. Also note from (3.6) that $\mathrm{Var}(Z) = 1$ and so ψ_0, ψ_p and ψ_* are given by

$$\psi_0(x) = \sigma_{12}x, \qquad \psi_p = \sigma_{13}x, \quad \psi_* = \sigma_{14}x.$$

We have thus verified (C4).

Now from (3.2) we have

THEOREM 3. *If the distribution of X_1 is nonarithmetic, $E(X_1^r) < \infty$ for some $r > 2$ and $E(Y_1)^4 < \infty$, then*

$$P\{A_2 \leq x\} = \Phi(x) + \frac{1}{\sqrt{a}}\lambda_2(x)\phi(x) + o_2\left(\frac{1}{\sqrt{a}}\right) \quad \text{as} \quad a \to \infty$$

where $\lambda_2(x) = (\sigma_{12} + 2\sigma_{13})(1 - x^2)/6 + \sigma_{14}x^2/2$, and

$$P\{A_1 \leq x\} = \Phi(x) + \frac{1}{\sqrt{a}}\lambda_1(x)\phi(x) + o_2\left(\frac{1}{\sqrt{a}}\right) \quad \text{as} \quad a \to \infty$$

where $\lambda_1(x) = (\sigma_{12} + 2\sigma_{13})(1 - x^2)/6$.

We still need to determine the parameters $\sigma_{12} + 2\sigma_{13}$ and σ_{14}, for which we have

THEOREM 4. *If the distribution of X_1 is nonarithmetic, $E(X_1^r) < \infty$ for some $r > 2$ and $E(Y_1)^6 < \infty$, then*

$$\sigma_{12} + 2\sigma_{13} = E(Y_1^3) - 3E(X_1Y_1) \quad \text{and} \quad \sigma_{14} = -E(X_1Y_1).$$

PROOF. Write (3.5) into matrix form: $\boldsymbol{l}^T B \boldsymbol{l} \to \boldsymbol{l}^T(\sigma_{ij})\boldsymbol{l}$ in probability where $B = (b_{ij})_{4\times 4}$ is symmetric and, in particular, has

$$b_{12} = \sum_{i=1}^{a} \frac{[U_i^2 - (\tau_i - \tau_{i-1}) - W_i]U_i}{a},$$

$$b_{13} = \sum_{i=1}^{a} \frac{[E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}}) - (\tau_i - \tau_{i-1}) - W_i]U_i}{a},$$

$$b_{14} = -\sum_{i=1}^{a} \frac{U_i W_i}{a}.$$

Write b_{14} as

$$b_{14} = -\sum_{i=1}^{a} \frac{U_i W_i - (\tau_i - \tau_{i-1})E(X_1'Y_1)}{a} - \frac{\tau_a E(X_1'Y_1)}{a}.$$

Then

$$\sum_{i=1}^{a} \frac{U_i W_i - (\tau_i - \tau_{i-1})E(X_1'Y_1)}{a} \xrightarrow{P} 0 \quad \text{as} \quad a \to \infty$$

since $\{U_i W_i - (\tau_i - \tau_{i-1})E(X_1'Y_1), i \geq 1\}$ are martingale increments by the second order Wald's lemma. Also note that $\tau_a/a \to 1$ a.s. So as $a \to \infty$

$$b_{14} \xrightarrow{P} \sigma_{14} = -E(X_1'Y_1) = -E(X_1Y_1).$$

Next we consider $b_{12} + 2b_{13}$:

$$\begin{aligned} b_{12} + 2b_{13} &= \sum_{i=1}^{a} \frac{U_i^3 - 3(\tau_i - \tau_{i-1})U_i - 3U_i W_i + 2U_i E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})}{a} \\ &= \sum_{i=1}^{a} \frac{U_i^3 - 3(\tau_i - \tau_{i-1})U_i - (\tau_i - \tau_{i-1})E(Y_1^3) + 2U_i E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}})}{a} \\ &\quad + \frac{\tau_a E(Y_1^3)}{a} + 3b_{14}. \end{aligned}$$

Now the first term above converges to zero in probability since $\{U_i^3 - 3(\tau_i - \tau_{i-1})U_i - (\tau_i - \tau_{i-1})E(Y_1^3), i \geq 1\}$ are martingale increments by the third order Wald's lemma (see e.g. Theorem 2.4.6 of Ghosh *et al.* (1997)) and $\{U_i E(U_i^2 \mid \mathcal{F}_{\tau_{i-1}}), i \geq 1\}$ are also martingale increments. Therefore as $a \to \infty$

$$b_{12} + 2b_{13} \xrightarrow{P} \sigma_{12} + 2\sigma_{13} = E(Y_1^3) - 3E(X_1Y_1).$$

This completes the proof.

4. A numerical example

Assume $(Z_1, V_1), (Z_2, V_2), \ldots$ are i.i.d. random vectors with $EZ_1 = \mu > 0$, $EV_1 = \nu$, $\mathrm{Var}(Z_1) = \sigma_1^2 > 0$, $\mathrm{Var}(V_1) = \sigma_2^2 > 0$ and $\mathrm{Cov}(Z_1, V_1) = \rho\sigma_1\sigma_2$. We are interested in the distribution of $\sqrt{\tau}(\bar{V}_\tau - \nu)/\sigma_2$, where τ is a stopping time given by

$$\tau = \inf\{n \geq 1 : Z_1 + \cdots + Z_n > a^*\}.$$

Table 1. Approximation $\Phi(x)$, our approximation (Appr) and the true value (Prob) of $P\{A_2 < x\}$.

		$\sigma_1 = 0.5$		$\sigma_1 = 1.0$		$\sigma_1 = 1.5$	
x	$\Phi(x)$	Appr	Prob	Appr	Prob	Appr	Prob
−1.8	0.0359	0.0322	0.0324	0.0297	0.0290	0.0293	0.0255
−1.5	0.0668	0.0614	0.0611	0.0567	0.0554	0.0552	0.0496
−1.2	0.1151	0.1089	0.1065	0.1015	0.0979	0.0984	0.0893
−0.9	0.1841	0.1733	0.1723	0.1642	0.1605	0.1590	0.1488
−0.6	0.2743	0.2625	0.2595	0.2510	0.2448	0.2436	0.2301
−0.3	0.3821	0.3685	0.3652	0.3552	0.3484	0.3456	0.3315
0.0	0.5000	0.4837	0.4824	0.4697	0.4647	0.4602	0.4471
0.3	0.6179	0.6060	0.6011	0.5931	0.5842	0.5800	0.5673
0.6	0.7257	0.7145	0.7110	0.7047	0.6963	0.6944	0.6816
0.9	0.8159	0.8083	0.8042	0.8018	0.7924	0.7910	0.7807
1.2	0.8849	0.8793	0.8763	0.8728	0.8678	0.8668	0.8592
1.5	0.9332	0.9304	0.9275	0.9267	0.9217	0.9216	0.9160
1.8	0.9641	0.9621	0.9606	0.9605	0.9571	0.9568	0.9536

Here Z and V may be the primary and a secondary variables, respectively, of interest in a study. While the stopping time τ of the study is determined by the primary variable Z, it is often necessary to make inference about the secondary variable V when the study stops and so the distribution of $\sqrt{\tau}(\bar{V}_\tau - \nu)/\sigma_2$ is of interest.

Using the result of Section 3, we can develop a simple approximation to the distribution function of $\sqrt{\tau}(\bar{V}_\tau - \nu)/\sigma_2$. Let $X_i = Z_i/\mu$ and $a = a^*/\mu$, then

$$\tau = \tau_a = \inf\{n \geq 1 : X_1 + \cdots + X_n > a\}.$$

Let $Y_i = (V_i - \nu)/\sigma_2$ then $\sqrt{\tau}(\bar{V}_\tau - \nu)/\sigma_2 = A_2$. From Theorems 3 and 4, $P\{A_2 \leq x\}$ can be approximated by

$$\Phi(x) + \frac{1}{\sqrt{a}}\{(\sigma_{12} + 2\sigma_{13})(1 - x^2)/6 + \sigma_{14}x^2/2\}\phi(x)$$

where

$$\sigma_{12} + 2\sigma_{13} = E(Y_1^3) - 3E(X_1Y_1) = E(V_1 - \nu)^3/\sigma_2^3 - 3\rho\sigma_1/\mu$$

and

$$\sigma_{14} = -E(X_1Y_1) = -\rho\sigma_1/\mu.$$

So

$$P\{A_2 \leq x\} \approx \Phi(x) + \frac{1}{\sqrt{a}}\left\{-\frac{\rho\sigma_1}{2\mu} + (1 - x^2)\frac{E(V_1 - \nu)^3}{6\sigma_2^3}\right\}\phi(x).$$

To assess the accuracy of this approximation, we did some numerical calculation for the special situation that (Z, V) is bivariate normal. In this case, the approximation becomes $\Phi(x) - \rho\sigma_1\phi(x)/(2\sqrt{a^*\mu})$. When $\mu = 1$ and $\sigma_2 = 1$, we calculated the simple normality approximation $\Phi(x)$, our approximation given above, and the true probability (which is estimated by Monte Carlo simulation with 100,000 replications), for selected values of x, σ_1, ρ and a^*. Table 1 contains the results for $\rho = 0.5$ (similar results are

observed for several other ρ values we tried) and for $a^* = 8.0$ (the results are more in favour of our approximation for $a^* > 8.0$).

It is clear from the table that our approximation is always better than the simple normality approximation, and the improvement becomes more substantial when σ_1 increases. Also, those Edgeworth-type approximations mentioned in Section 1 are not applicable to this setting.

Acknowledgements

We thank Professor Woodroofe for inspiring us to do this research, and the referees for many helpful comments.

References

Asmussen, S. (1987). *Applied Probability and Queues*, Wiley, New York.

Bose, A. and Boukai, B. (1996). Estimation with prescribed proportional accuracy for a two-parameter exponential family of distributions, *Ann. Statist.*, **24**, 1792–1803.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory*, Springer, New York.

Chow, Y. S. and Yu, K. F. (1981). The performance of a sequential procedure for the estimation of the mean, *Ann. Statist.*, **9**, 184–189.

Ghosh, M., Mukhopadhyay, N. and Sen, P. K. (1997). *Sequential Estimation*, Wiley, Now York.

Gut, A. (1988). *Stopped Random Walks*, Springer, New York.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

Lai, T.L. and Wang, J.Q. (1994). Asymptotic expansions for the distributions of stopped random walks and first passage times, *Ann. Probab.*, **22**, 269–284.

Landers, D. and Rogge, L. (1976). The exact approximation order in the central limit theorem for random summation, *Z. Wahrsch. Verw. Gebiete.*, **36**, 269–284.

Mykland, P. A. (1992). Asymptotic expansions and bootstrapping-distributions for dependent variables: A martingale approach, *Ann. Statist.*, **20**, 623–654.

Mykland, P. A. (1993). Asymptotic expansions for martingales, *Ann. Probab.*, **21**, 800–818.

Rao, M. M. (1984). *Probability Theory with Applications*, Academic Press, New York.

Rényi, A. (1957). On the asymptotic distribution of the sum of a random number of independent random variables, *Acta Mathematica Academiae Scientiarum Hungaricae*, **8**, 193–199.

Revuz, D. (1975). *Markov Chains*, North-Holland, Amsterdam.

Takahashi, H. (1987). Asymptotic expansions in Anscombe's theorem for repeated significance tests and estimation after sequential testing, *Ann. Statist.*, **15**, 278–295.

Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, SIAM, Philadelphia, Pennsylvania.

Woodroofe, M. and Keener, R. (1987). Asymptotic expansion for boundary crossing problem, *Ann. Probab.*, **15**, 102–114.