Elect. Comm. in Probab. **7** (2002) 47–54

ELECTRONIC COMMUNICATIONS in PROBABILITY

ASYMPTOTICS FOR PRODUCTS OF SUMS AND U-STATISTICS

GRZEGORZ REMPAŁA[1]
Department of Mathematics, University of Louisville, Louisville, 40241 USA
email: `grzes@louisville.edu`

JACEK WESOŁOWSKI[2]
Wydzial Matematyki i Nauk Informacyjnych, Politechnika Warszawska, Pl. Politechniki 1, 00-661 Warszawa, Poland
email: `wesolo@alpha.mini.pw.edu.pl`

submitted January 15, 2002 *Final version accepted* January 20, 2002

AMS 2000 Subject classification: 60F05, 62G32
central limit theorem, lognormal distribution, products of sums of iid rv's, records, U-statistics

Abstract

The product of subsequent partial sums of independent, identically distributed, square integrable, positive random variables is asymptotically lognormal. The result extends in a rather routine way to non-degenerate U-statistics.

1 Introduction

It is well known that the products of independent, identically distributed (iid), positive, square integrable random variables (rv's) are asymptotically lognormal. This fact is an immediate consequence of the classical central limit theorem (clt). In the present paper, we are interested in the limiting law of the products of sums of iid rv's. It appears that their asymptotic behavior is rather similar. We also derive a parallel result for the products of non-degenerate U-statistics.

While considering limiting properties of sums of records, Arnold and Villaseñor (1998) obtained the following version of the clt for a sequence (X_n) of iid exponential rv's with the mean equal one

$$\frac{\sum_{k=1}^{n}\log(S_k) - n\log(n) + n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N} \tag{1}$$

as $n \to \infty$, where $S_k = X_1 + \ldots + X_k$, $k = 1, 2, \ldots$, and $\mathcal{N}$ is a standard normal rv. Their proof is heavily based on a very special property of exponential(gamma) distributions: namely that there is independence of ratios of subsequent partial sums and the last sum. It uses also Resnick's (1973) result on weak limits for records.

Observe that, via the Stirling formula, the relation (1) can be equivalently stated as

$$\left(\prod_{k=1}^{n} \frac{S_k}{k}\right)^{\frac{1}{\sqrt{n}}} \xrightarrow{d} e^{\sqrt{2}\mathcal{N}} .$$

The main purpose of this note is to show that this limit behavior of a product of partial sums has a universal character and holds true for any sequence of square integrable positive iid rv's. This is done in Section 2. Section 3 extends the result to a U-statistics setting.

2 Main Result

Our main result is a general version of (1), without assuming any particular distribution for the X_i's.

Theorem 1. *Let (X_n) be a sequence of iid positive square integrable rv's. Denote $\mu = E(X_1) > 0$, the coefficient of variation $\gamma = \sigma/\mu$, where $\sigma^2 = Var(X_1)$, and $S_k = X_1 + \ldots + X_k$, $k = 1, 2, \ldots$. Then*

$$\left(\frac{\prod_{k=1}^{n} S_k}{n!\mu^n}\right)^{\frac{1}{\gamma\sqrt{n}}} \xrightarrow{d} e^{\sqrt{2}\mathcal{N}} , \tag{2}$$

where $\mathcal{N}$ is a standard normal rv.

Before proving the above result we will establish a version of the classical central limit theorem, essentially, for scaled iid rv's. To this end we will use the clt for triangular arrays, so the basic step in the proof will rely on verifying the Lindeberg condition.

Lemma 1. *Under the assumptions of Theorem 1*

$$\frac{1}{\gamma\sqrt{2n}} \sum_{k=1}^{n} \left(\frac{S_k}{\mu k} - 1\right) \xrightarrow{d} \mathcal{N} . \tag{3}$$

Proof of Lemma 1. Let $Y_i = (X_i - \mu)/\sigma$, $i = 1, 2, \ldots$, and denote $\tilde{S}_k = Y_1 + \ldots + Y_k$, $k = 1, 2, \ldots$. Then (3) becomes

$$\frac{1}{\sqrt{2n}} \sum_{k=1}^{n} \frac{\tilde{S}_k}{k} \xrightarrow{d} \mathcal{N} .$$

Observe that

$$\sum_{k=1}^{n} \frac{\tilde{S}_k}{k} = \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{k} Y_i = \sum_{i=1}^{n} b_{i,n} Y_i ,$$

where

$$b_{i,n} = \sum_{k=i}^{n} \frac{1}{k} , \quad i = 1, \ldots, n.$$

Define now

$$Z_{i,n} = \frac{b_{i,n}}{\sqrt{2n}} Y_i ,$$

and observe that $E(Z_{i,n}) = 0$ and

$$Var(Z_{i,n}) = \frac{b_{i,n}^2}{2n} \ , \quad i = 1, \ldots, n.$$

Also, since for $k \geq l$,

$$Cov\left(\frac{\tilde{S}_k}{k}, \frac{\tilde{S}_l}{l}\right) = \frac{1}{k},$$

then

$$Var\left(\sum_{i=1}^n Z_{i,n}\right) = \frac{1}{2n} Var\left(\sum_{k=1}^n \frac{\tilde{S}_k}{k}\right) = \frac{1}{2n}\left(b_{1,n} + 2\sum_{k=2}^n \sum_{l=1}^{k-1} \frac{1}{k}\right) =$$

$$= \frac{1}{2n}\left(b_{1,n} + 2\sum_{k=2}^n \frac{k-1}{k}\right) = 1 - \frac{b_{1,n}}{2n} \to 1$$

as $n \to \infty$. Observe that as a by-product of the above computation we have obtained the identity

$$\sum_{k=1}^n b_{i,n}^2 = 2n - b_{1,n} \ . \tag{4}$$

In order to complete the proof we need to check that the Lindeberg condition is satisfied for the triangular array $[Z_{i,n}]$. Take any $\varepsilon > 0$. Then, using (4), we get

$$\sum_{i=1}^n E(Z_{i,n}^2 I(|Z_{i,n}| > \varepsilon)) = \frac{1}{2n}\sum_{i=1}^n b_{i,n}^2 E\left(Y_i^2 I\left(|Y_i| > \frac{\varepsilon\sqrt{2n}}{b_{i,n}}\right)\right) \leq$$

$$\leq \frac{1}{2n}\sum_{i=1}^n b_{i,n}^2 E\left(Y_i^2 I\left(|Y_i| > \frac{\varepsilon\sqrt{2n}}{\log(n)}\right)\right) = \frac{a_n}{2n}\sum_{i=1}^n b_{i,n}^2 = a_n\left(1 - \frac{b_{1,n}}{2n}\right) \ ,$$

where

$$a_n = E\left(Y_i^2 I\left(|Y_i| > \frac{\varepsilon\sqrt{2n}}{\log(n)}\right)\right)$$

does not depend on i. Since $a_n \to 0$ as $n \to \infty$ then the Lindeberg condition holds. $\square$

Proof of Theorem 1. The proof relies on the delta-method expansion. In what follows we use only elementary considerations to justify its validity.
Denote $C_k = S_k/(\mu k)$, $k = 1, 2, \ldots$. By the strong law of large numbers it follows that for any $\delta > 0 \; \exists \, R$ such that $\forall \, r > R$

$$P(\sup_{k \geq r} |C_k - 1| > \delta) < \delta.$$

Consequently, there exist two sequences $(\delta_m) \downarrow 0$ $(\delta_1 = 1/2)$ and $(R_m) \uparrow \infty$ such that

$$P(\sup_{k \geq R_m} |C_k - 1| > \delta_m) < \delta_m.$$

Take now any real x and any m. Then

$$P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^n \log(C_k) \leq x\right) = P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^n \log(C_k) \leq x \ , \ \sup_{k > R_m} |C_k - 1| > \delta_m\right) +$$

$$+P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{n}\log(C_k)\le x\ ,\ \sup_{k>R_m}|C_k-1|\le\delta_m\right)=A_{m,n}+B_{m,n}$$

and $A_{m,n}<\delta_m$.
To compute $B_{m,n}$ we will expand the logarithm: $\log(1+x)=x+\frac{x^2}{(1+\theta x)^2}$, where $\theta\in(0,1)$ depends on $x\in(-1,1)$. Thus,

$$B_{m,n}=P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}\log(C_k)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}\log(1+(C_k-1))\le x\ ,\ \sup_{k>R_m}|C_k-1|\le\delta_m\right)=$$

$$=P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}\log(C_k)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}(C_k-1)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}\frac{(C_k-1)^2}{(1+\theta_k(C_k-1))^2}\le x\ ,\right.$$

$$\left.\sup_{k>R_m}|C_k-1|\le\delta_m\right)=P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}\log(C_k)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}(C_k-1)+\right.$$

$$\left.\left[\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}\frac{(C_k-1)^2}{(1+\theta_k(C_k-1))^2}\right]I(\sup_{k>R_m}|C_k-1|\le\delta_m)\le x\right)-$$

$$-P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}\log(C_k)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}(C_k-1)\le x,\ \sup_{k>R_m}|C_k-1|>\delta_m\right)=D_{m,n}+F_{m,n}.$$

where θ_k, $k=1,\ldots,n$ are $(0,1)$-valued rv's and $F_{m,n}<\delta_m$.
Rewrite now $D_{m,n}$ as

$$D_{m,n}=P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}(\log(C_k)-C_k+1)+\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{n}(C_k-1)+\right.$$

$$\left.+\left[\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}\frac{(C_k-1)^2}{(1+\theta_k(C_k-1))^2}\right]I(\sup_{k>R_m}|C_k-1|<\delta_m)\le x\right).$$

Observe now that for any fixed m

$$\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{R_m}(\log(C_k)-C_k+1)\xrightarrow{P}0 \tag{5}$$

as $n\to\infty$ (as a matter of fact this sequence converges to zero a.s.).
Note that for $|x|<1/2$ and any $\theta\in(0,1)$ it follows that $x^2/(1+\theta x)^2\le 4x^2$. Then for any m

$$\left[\frac{1}{\gamma\sqrt{2n}}\sum_{k=R_m+1}^{n}\frac{(C_k-1)^2}{(1+\theta_k(C_k-1))^2}\right]I(\sup_{k>R_m}|C_k-1|<\delta_m)\le\frac{4}{\sqrt{2n}}\sum_{k=1}^{n}(C_k-1)^2\xrightarrow{P}0, \tag{6}$$

as $n\to\infty$. The above is a consequence of the Markov inequality, since for any $\varepsilon>0$

$$P\left(\frac{4}{\sqrt{2n}}\sum_{k=1}^{n}(C_k-1)^2>\varepsilon\right)\le\frac{4}{\varepsilon\sqrt{2n}}\sum_{k=1}^{n}Var(C_k)=\frac{4}{\varepsilon\sqrt{2n}}\sum_{k=1}^{n}\frac{1}{k}\to 0$$

as $n \to \infty$ for any fixed m.
Since by Lemma 1 it follows that

$$\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{n}(C_k - 1) \xrightarrow{d} N$$

as $n \to \infty$ then by (5) and (6) we conclude that for any fixed m

$$D_{m,n} \to \Phi(x) ,$$

where Φ is the standard normal distribution function.
Finally, observe that

$$P\left(\log\left(\frac{\prod_{k=1}^{n} S_k}{n!\mu^n}\right)^{\frac{1}{\gamma\sqrt{n}}} \le x\right) = P\left(\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{n}\log(C_k) \le x\right) = A_{m,n} + D_{m,n} + F_{m,n}$$

which implies (2) since $A_{m,n} + F_{m,n} < 2\delta_m \to 0$ as $m \to \infty$, uniformly in n. □

Remark 1. It is perhaps worth to notice that by the strong law of large numbers and the property of the geometric mean it follows directly that

$$\left(\frac{\prod_{k=1}^{n} S_k}{n!}\right)^{\frac{1}{n}} \to \mu \quad a.s.$$

if only existence of the first moment is assumed.

Remark 2. In Arnold and Villaseñor (1998) the following identity was proved:

$$T_n = \sum_{k=1}^{n}\log(\sum_{i=1}^{k} X_k) \stackrel{d}{=} -\sum_{i=1}^{n}\tilde{X}_i + nR_n$$

where (X_n) and $(\tilde{X}_n)$ are independent sequences of iid standard exponential rv's and R_n is nth record from an independent sequence of iid Gumbel rv's (R_1 is the first observation). Consequently

$$\frac{T_n - n\log(n) + n}{\sqrt{2n}} \stackrel{d}{=} \frac{1}{\sqrt{2}}\left(-\frac{\sum_{i=1}^{n}\tilde{X}_i - n}{\sqrt{n}} + \sqrt{n}(R_n - \log(n))\right) .$$

Now, in view of the result above, we can apply the argument used in Arnold and Villaseñor (1998) in the reverse order. From Theorem 1 it follows that the left hand side converges in distribution to the standard normal law and the same holds true by the standard clt for the first element at the rhs. Now since the first and the second element at the rhs are independent it follows that $\sqrt{n}(R_n - \log(n))$ is asymptotically standard normal, which proves Resnick's result for records from the Gumbel distribution.

3 Extension to U-statistics

A useful notion of a "U-statistic" has been introduced by Hoeffding (1948) and pertains to any estimator based on the statistic of the form

$$U_n = \sum_{1\le i_1<\ldots<i_m\le n} h(X_{i_1},\ldots,X_{i_m}) \tag{7}$$

where h is a symmetric real function of m arguments, the X_i's are iid rv's, and the summation is carried over all possible choices of m indices out of the set $\{1, 2, \ldots, n\}$. Let us note that if $m = 1$ and $h(x) = x$ then the above definition gives simply S_n. If we assume that $E\, h(X_1, \ldots, X_m)^2 < \infty$ and define $h_1(x) = E\, h(x, X_2, \ldots, X_m)$ as well as

$$\hat{U}_n = \binom{n}{m} \left[\frac{m}{n} \sum_{i=1}^{n} (h_1(X_i) - Eh) + E\, h \right],$$

then we may write

$$U_n = \hat{U}_n + R_n \tag{8}$$

where

$$R_n = \sum_{1 \le i_1 < \ldots < i_m \le n} H(X_{i_1}, \ldots, X_{i_m}),$$

and

$$H(x_1, \ldots, x_m) = h(x_1, \ldots, x_m) - \sum_{i=1}^{m} (h_1(x_i) - E\, h) - E\, h.$$

It is well known (cf. e.g., Serfling 1980) that,

$$Cov(\hat{U}_n, R_n) = 0$$

and

$$n\, Var\left[\binom{n}{m}^{-1} R_n \right] \to 0 \quad \text{as} \quad n \to \infty. \tag{9}$$

The result of Theorem 1 can be extended to U-statistics as follows.

Theorem 2. *Let U_n be a statistic given by (7). Assume $E\, h^2 < \infty$ and $P(h(X_1, \ldots, X_m) > 0) = 1$, as well as $\sigma^2 = Var(h_1(X_1)) \neq 0$. Denote $\mu = E\, h > 0$ and $\gamma = \sigma/\mu > 0$, the coefficient of variation. Then*

$$\left(\prod_{k=m}^{n} \frac{U_k}{\binom{k}{m} \mu} \right)^{\frac{1}{m\gamma\sqrt{n}}} \xrightarrow{d} e^{\sqrt{2}\mathcal{N}},$$

where $\mathcal{N}$ is a standard normal rv.

In order to prove the theorem we shall first consider a more general version of (3).

Lemma 2. *Under the assumptions of Theorem 2*

$$\frac{1}{m\, \gamma \sqrt{2n}} \sum_{k=m}^{n} \left(\frac{U_k}{\mu \binom{k}{m}} - 1 \right) \xrightarrow{d} \mathcal{N}.$$

Proof of Lemma 2. Using the decomposition (8) we have

$$\frac{1}{m\, \gamma \sqrt{2n}} \sum_{k=m}^{n} \left(\frac{U_k}{\mu \binom{k}{m}} - 1 \right) = \frac{1}{m\, \gamma \sqrt{2n}} \sum_{k=m}^{n} \left(\frac{\hat{U}_k}{\mu \binom{k}{m}} - 1 \right) + \frac{1}{m\, \sigma \sqrt{2n}} \sum_{k=m}^{n} \frac{R_k}{\binom{k}{m}}.$$

By Lemma 1, applied to the rv's $m\,h_1(X_i)$ for $i = 1, 2, \dots$ we have

$$\frac{1}{m\,\gamma\sqrt{2n}}\sum_{k=m}^{n}\left(\frac{\hat{U}_k}{\mu\binom{k}{m}}-1\right)=\frac{1}{\gamma\sqrt{2n}}\sum_{k=1}^{n}\left(\frac{\sum_{i=1}^{k}h_1(X_i)}{\mu k}-1\right)-\sum_{k=1}^{m-1}\left(\frac{\sum_{i=1}^{k}h_1(X_i)}{\mu k}-1\right)\xrightarrow{d}\mathcal{N}$$

since the second expression converges to zero a.s. as $n \to \infty$. Therefore, in order to prove the lemma it suffices to show

$$\tilde{R}_n = \frac{1}{m\,\sigma\sqrt{2n}}\sum_{k=m}^{n}\frac{R_k}{\binom{k}{m}} \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty.$$

To argue the above, it is, in turn, enough to argue that

$$E\,\tilde{R}_n^2 \to 0 \quad \text{as} \quad n \to \infty.$$

To this end let us note that, due to symmetries involved, we have

$$Cov\left[\frac{R_l}{\binom{l}{m}}, \frac{R_k}{\binom{k}{m}}\right] = Var\left[\frac{R_k}{\binom{k}{m}}\right] \quad \text{for} \;\; l < k,$$

and thus

$$\begin{aligned}
E\,\tilde{R}_n^2 = Var\,\tilde{R}_n &= \frac{1}{m^2\,\sigma^2\,2n}\,Var\left[\sum_{k=m}^{n}\frac{R_k}{\binom{k}{m}}\right]\\
&= \frac{1}{m^2\,\sigma^2\,2n}\left\{\sum_{k=m}^{n}Var\left[\frac{R_k}{\binom{k}{m}}\right] + 2\sum_{m\le l<k\le n}Cov\left[\frac{R_k}{\binom{k}{m}}, \frac{R_l}{\binom{l}{m}}\right]\right\}\\
&= \frac{1}{m^2\,\sigma^2\,2n}\left\{\sum_{k=m}^{n}Var\left[\frac{R_k}{\binom{k}{m}}\right] + 2\sum_{m\le l<k\le n}Var\left[\frac{R_k}{\binom{k}{m}}\right]\right\}\\
&= \frac{1}{m^2\,\sigma^2\,2n}\left\{\sum_{k=m}^{n}(1+2k-2m)\,Var\left[\frac{R_k}{\binom{k}{m}}\right]\right\} \to 0
\end{aligned}$$

as $n \to \infty$, in view of (9) and the property of arithmetic mean. □

Proof of Theorem 2. In view of the fact that if $E\,|h| < \infty$ then $\binom{n}{m}^{-1}U_n \to E\,h = \mu \;\; a.s.$ (see, e.g., Serfling 1980) and Lemma 2 above, the argument used in the proof of Theorem 1 can be virtually repeated with S_k/k replaced now by $\binom{k}{m}^{-1}U_k$ and γ by $m\gamma$. □

Remark 3. Let us note that in view of the strong law of large numbers for U-statistics we may extend the observation of Remark 1 to U-statistics as follows

$$\left(\prod_{k=m}^{n}\frac{U_k}{\binom{k}{m}}\right)^{\frac{1}{n}} \to \mu \quad a.s.$$

if $E\,|h| < \infty$.

References

Arnold, B.C., Villaseñor, J.A. (1998) The asymptotic distribution of sums of records. *Extremes* **1:3**, 351-363.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons Inc., New York.

Resnick, S.I. (1973) Limit laws for record values. *Stochastic Processes and Their Applications* **1**, 67-82.