Statistica Sinica **19** (2009), 315-323

ASYMPTOTIC OVERSHOOTS FOR ARITHMETIC I.I.D. RANDOM VARIABLES

I-Ping Tu

Academia Sinica, Taipei

Abstract: The Sequential Probability Ratio Test (SPRT) has been widely applied in quality control and clinical studies. There are two important quantities in SPRT: $[1 - E(e^{-\alpha S_{\tau_+}})]/E(S_{\tau_+})$ for calculating the p-value and $E(S^2_{\tau_+})/2E(S_{\tau_+})$ for estimating the sample size, where S_n is the i.i.d. summation of random variables and τ_+ refers to the first time that S_n becomes positive. For non-arithmetic i.i.d. random variables, Woodroofe (1979) provided computation formulas for these two quantities. To find the threshold for the IBD score statistics in testing genetic linkage, Tu and Siegmund (1999) provided a computation formula to calculate $[1 - E(e^{-\alpha S_{\tau_+}})]/(1 - e^{-\alpha h})E(S_{\tau_+})$ for arithmetic i.i.d. random variables when α is not too small. This paper gives another computation formula to calculate $[1 - E(e^{-\alpha S_{\tau_+}})]/(1 - e^{-\alpha h})E(S_{\tau_+})$ for arithmetic i.i.d. random variables, which can be applied for any positive α including $\alpha \downarrow 0$. We also provide a computation formula for $E(S^2_{\tau_+})/2E(S_{\tau_+})$ to estimate the overshoot for arithmetic i.i.d. random variables. Furthermore, we show that these two formula reproduce Woodroofe's non-arithmetic formula by letting the span h go to zero, and we derive a computation formula to calculate $E(S_{\tau_+})$, that can be applied to estimate the number of 'new-high' points in reaching a threshold.

Key words and phrases: Arithmetic, ladder height, overshoot, sample size, second moment, sequential analysis.

1. Introduction

A random variable x is called arithmetic if it takes only values 0, $\pm h$, $\pm 2h$, $\ldots$, and h is called its span if h is the largest number that satisfies $P(x \in \{0, \pm h, \pm 2h, \ldots\}) = 1$. Let x_1, $x_2 \ldots$ be i.i.d. from f_ω with positive mean μ, and consider a test that terminates at $n = T$ when T is the first time that $S_n = \sum_{i=1}^n x_i$ exceeds a threshold b, $T = \inf\{n : S_n \geq b\}$. Two basic problems encountered in sequential tests are the p-value calculation under the null hypothesis, to find an appropriate threshold b, and the sample size estimation under the alternative hypothesis, to estimate cost.

S_T can be rewritten as an independent sum of positive random variables S_{τ_+}. Let $\tau_+^{(0)} = 0$, $\tau_+ = \tau_+^{(1)} = \inf\{n : S_n > 0\}$, $\tau_+^{(2)} = \inf\{n : S_n > S_{\tau_+^{(1)}}\}$, $\ldots$, $\tau_+^{(k)} = \inf\{n : S_n > S_{\tau_+^{(k-1)}}\}$, and observe that $\sum_{k=1}^\infty P(T = \tau_+^{(k)} | T < \infty) = 1$,

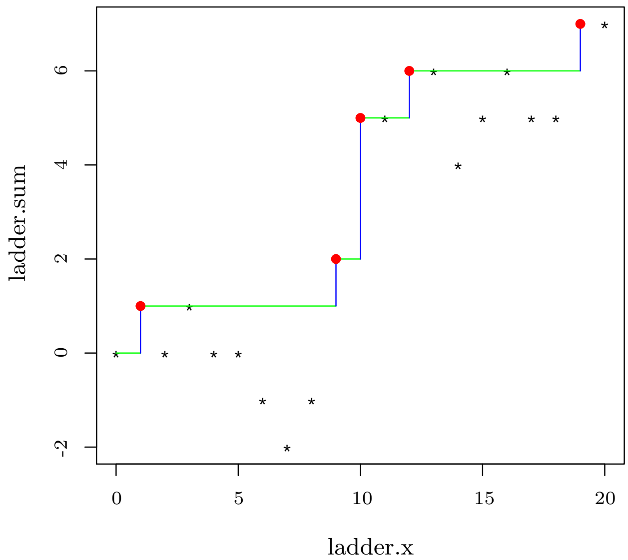

Figure 1. The solid red circles are examples of new-high points $(\tau_+^{(i)}, S_{\tau_+^{(i)}})$, where $i = 1, \ldots, 5$. The green (horizontal) lines describe the waiting time to reach the new-high points, and the blue (vertical) lines are the jump sizes of the new-high points. It can be observed from the figure that the first time to exceed some positive threshold must occur at a new-high point $\tau_+^{(i)}$ for some $i > 0$, that is, $S_T = S_{\tau_+^{(i)}}$ for some $i > 0$.

and $S_{\tau_+^{(1)}}, S_{\tau_+^{(2)}} - S_{\tau_+^{(1)}}, S_{\tau_+^{(3)}} - S_{\tau_+^{(2)}}, \ldots$ are positive with probability 1 and i.i.d.. This relates S_T to S_{τ_+}. Figure 1 gives an illustration of these ladder height random variables. We call the points $(\tau_+^{(k)}, S_{\tau_+^{(k)}})$ 'new-high' points.

For a negative drift process, by applying the change of measure method, one can write

$$P(T < \infty) = \sum_{n=1}^{\infty} E_0(I_{[T=n]}) = \sum_{n=1}^{\infty} E_\alpha(\exp(-S_n\alpha + n\varphi(\alpha))I_{[T=n]})$$
$$= e^{(-\alpha b)} E_\alpha(e^{-\alpha(S_T - b)}) \tag{1.1}$$

where $\varphi(\alpha)$ is the log of the moment generating function of x_1, with α referring to the change of measure chosen by solving the equation $\varphi(\alpha) = 0$ under the constraint that the mean, $\varphi'(\alpha)$, is positive. In (1.1), $E_\alpha(e^{-\alpha(S_T-b)})$ can be treated as a correction term for the contribution of the overshoot. Siegmund (1985, Chap. 8) provided asymptotic equations for the correction term in (1.1):

for i.i.d. non-arithmetic random variables,

$$\lim_{b\to\infty} E_\alpha(e^{-\alpha(S_T-b)}) = \frac{1-E_\alpha(e^{-\alpha S_{\tau_+}})}{\alpha E_\alpha(S_{\tau_+})}; \tag{1.2}$$

for i.i.d. arithmetic random variables

$$\lim_{b\to\infty} E_\alpha(e^{-\alpha(S_T-b)}) = h\frac{1-E_\alpha(e^{-\alpha S_{\tau_+}})}{(1-e^{-\alpha h})E_\alpha(S_{\tau_+})}. \tag{1.3}$$

Woodroofe (1979) provided a computation formula for (1.2), namely

$$\begin{aligned}\frac{1-E(\exp(-\alpha S_{\tau_+}))}{\alpha E(S_{\tau_+})} \\ = \exp\Big\{\frac{1}{\pi}\int_0^\infty \Big[\frac{\alpha^2}{\alpha^2+t^2}\frac{\Im(\xi(t))-\frac{\pi}{2}}{t} - \frac{\alpha}{\alpha^2+t^2}(\Re(\xi(t))+\log(\mu t))\Big]dt\Big\},\end{aligned} \tag{1.4}$$

while Tu and Siegmund (1999) provided a computation formula for (1.3):

$$\begin{aligned}h\frac{1-E(\exp(-\alpha S_{\tau_+}))}{(1-\exp(-\alpha h))E(S_{\tau_+})} = \frac{1}{(1-\exp(-\alpha h))}\exp\{\frac{-1}{2\pi}\int_0^{2\pi} dt[\frac{\xi(t/h)\exp(-\alpha h-it)}{1-\exp(-\alpha h-it)} \\ +\frac{\xi(t/h)+\log(\mu(1-\exp(it))/h)}{1-\exp(it)}]\}.\end{aligned} \tag{1.5}$$

(1.5) has been applied to approximate the tail probability of IBD scores in genetic linkage problems in Tu and Siegmund (1999), the scan statistics for arithmetic cases on genomic sequence alignment in Storey and Siegmund (2001), and the weighted scores of specific patterns on genomic sequence in Chan and Zhang (2007). (1.5) is correct for α not too close to 0. However, when $\alpha \downarrow 0$, it may break down. In Theorem 1, we rewrite (1.5) and show that the new formula can reproduce (1.4) in Woodroofe (1979) by letting $h \downarrow 0$ in Corollary 1.

Under the alternative hypothesis, where $0 < E(x_1) < \infty$,

$$E(T) = \frac{b}{\mu} + \frac{E(S_T-b)}{\mu}. \tag{1.6}$$

In estimating the sample size, $E(S_T-b)/\mu$ can be viewed as a correction term for the overshoot. With the condition $E(x_1^2) < \infty$, Siegmund (1985) provided the asymptotic equations for this correction term: for i.i.d. non-arithmetic random variables,

$$\lim_{b\to\infty} E(S_T-b) = \frac{E(S_{\tau_+}^2)}{2E(S_{\tau_+})}; \tag{1.7}$$

for i.i.d. arithmetic random variables,

$$\lim_{b\to\infty} E(S_T-b) = \frac{E(S_{\tau_+}^2)}{2E(S_{\tau_+})} - \frac{h}{2}. \tag{1.8}$$

For (1.7), Woodroofe (1979) showed that

$$\frac{E(S_{\tau_+}^2)}{2E(S_{\tau_+})} = \frac{E(x_1^2)}{4E(x_1)} + \frac{1}{\pi}\int_0^\infty \frac{(\Re(\xi(t)) + \log(\mu t))}{t^2}dt. \tag{1.9}$$

In this paper, we provide an expression for (1.8) in Theorem 2, and prove that it converges to (1.9) as $h \downarrow 0$.

This paper is organized as follows. The main results are in Section 2 and are illustrated by several examples in Section 3. Results are applied to estimate the overshoot, the waiting time and the number of 'new-high' points, and then compared with simulations. The paper ends with a discussion section. The technical proofs are put in an appendix that is available at (http://www3.stat.sinica.edu.tw /statistica/).

2. Main Results

Let $x_1, x_2, \ldots$ be arithmetic i.i.d. random variables with span h, $\mu = E(x_1) > 0$, $E(x_1^2) < \infty$, and $\tau_+ = \inf\{n : S_n > 0\}$. Let $\phi(t) = E(\exp(itx_1))$, $\xi(t) = \sum_{n=1}^\infty \phi^n(t)/n = -\log(1-\phi(t))$, $\Re(\xi(t)) = -(1/2)\log((1-\Re(\phi(t)))^2 + (\Im(\phi(t)))^2)$, and $\Im(\xi(t)) = \tan^{-1}[\Im(\phi(t))/(1-\Re(\phi(t)))]$. ($\Re$ means real part and $\Im$ means imaginary part of complex variables).

Theorem 1. In the given notation,

$$h\frac{1 - E(\exp(-\alpha S_{\tau_+}))}{(1-\exp(-\alpha h))E(S_{\tau_+})} = \exp\Big\{\frac{-h}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt\Big[\Big(\xi(t)+\log(\frac{\mu(1-e^{ith})}{h})\Big)\Big(\frac{e^{-\alpha h - iht}}{1-e^{-\alpha h-iht}} + \frac{1}{1-e^{iht}}\Big)\Big]\Big\}. \tag{2.1}$$

Corollary 1. With the condition $\limsup|\phi(t)| < 1$, (1.4) in Woodroofe (1979) can be reproduced by taking the limit as $h \downarrow 0$ in (2.1):

$$\lim_{h\downarrow 0}\frac{-h}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt\Big[\Big(\xi(t) + \log(\frac{\mu(1-e^{ith})}{h})\Big)\Big(\frac{e^{-\alpha h-iht}}{1-e^{-\alpha h-iht}} + \frac{1}{1-e^{iht}}\Big)\Big]$$
$$= \frac{1}{\pi}\int_0^\infty \Big[\frac{\alpha^2}{\alpha^2+t^2}\frac{\Im(\xi(t)) - \frac{\pi}{2}}{t} - \frac{\alpha}{\alpha^2+t^2}(\Re(\xi(t)) + \log(\mu t))\Big]dt.$$

Theorem 2. One has

$$\frac{E(S_{\tau_+}^2)}{2E(S_{\tau_+})} = \frac{E(x_1^2)}{4\mu} + \frac{h}{4} - \frac{h^2}{4\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt\frac{[\Re(\xi(t)) + \log(\frac{\mu}{h}) + \log(2|\sin(\frac{ht}{2})|)]}{\cos(ht)-1}. \tag{2.2}$$

Corollary 2. With the condition $\limsup |\phi(t)| < 1$, (1.9) in Woodroofe (1979) can be reproduced by taking the limit as $h \downarrow 0$ in (2.2):

$$\lim_{h\downarrow 0} \frac{h}{4} - \frac{h^2}{4\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt \frac{[\Re(\xi(t)) + \log(\frac{\mu}{h}) + \log(2|\sin(\frac{ht}{2})|)]}{\cos(ht) - 1}$$
$$= \frac{1}{\pi}\int_0^\infty \frac{(\Re(\xi(t)) + \log(\mu t))}{t^2} dt$$

Theorem 3. *One has*

$$\log[\frac{E(S_{\tau_+})}{h}] = \frac{h}{4\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt \left(\Re(\xi(t)) + \log(\frac{\mu}{h}) + \log(2|\sin(\frac{ht}{2})|)\right)$$
$$-\frac{h}{4\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} dt \left([\Im(\xi(t)) - \frac{\pi}{2} + \frac{ht}{2}]\frac{1}{\tan(\frac{ht}{2})}\right). \quad (2.3)$$

3. Examples and Simulations

We give some examples to demonstrate how the theorems work.

3.1. Example 1: Bernoulli random variables

Bernoulli random variable with parameter $p > 0$ is a place to start because we know that $S_{\tau_+} = 1$ with probability 1, which means that (2.1) should output 1, (2.2) should output .5, and (2.3) should output 0. The characteristic function for a Bernoulli random variable is $\phi(t) = (1-p) + p\exp(it)$, and $\xi(t) = -\log(1-\phi(t)) = -\log(p(1-e^{it}))$, such that $\xi(t) + \log(\mu(1-e^{it})) = 0$ so (2.1) outputs 0 as expected. The real part of $\xi(t)$ is $\Re(\xi(t)) = -\log(p) - \log[\sqrt{(1-\cos(t))^2 + \sin^2 t} = -\log(p) - \log(2|\sin(t/2)|)$, and the imaginary part for $\xi(t)$ is $\Im(\xi(t)) = \tan^{-1}(\cot(t/2)) = \pi/2 - t/2$. Because $E(x_1) = E(x_1^2) = p$ and the span $h = 1$, the three integrals in (2.2) and (2.3) are 0 with $E(S_{\tau_+}^2)/2E(S_{\tau_+}) = 0.5$ and $\log[E(S_{\tau_+})] = 0$, as expected.

3.2. Example 2: Mixture of Poisson random variables

This example comes from the conditional behavior of a function of a Markov Chain (Tu and Siegmund (1999)). Let $\{x_i, i \geq 1\}$ be i.i.d. with $x_1 = 3y_1 + y_2 - y_3 - 3y_4$, where the y_i's are i.i.d. Poisson variables with parameters λ_i. We adjust the parameters of the Poisson random variables linearly in such a way that $\lambda_i = sp_i$, where $\sum_{i=1}^4 p_i = 1$, and $0 < s < \infty$. When $s \to 0$, S_{τ_+} will be very similar to $S_{\tau_+}^{(r)}$ of a random walk with jump sizes 3, 1, -1, -3, and with probabilities p_1, p_2, p_3 and p_4. $S_{\tau_+}^{(r)}$ can be solved exactly, and the solution can provide a check.

Table 1. Numerical calculations of $[1-E(e^{-\alpha S^{(r)}_{\tau+}})]/(1-e^{-\alpha})E(S^{(r)}_{\tau+})$ for the random walk $(3,1,-1,-3)$ with $p_1 = 0.6157$, $p_2 = 0.0269$, $p_3 = 0.3304$ and $p_4 = 0.0269$ are shown.

Calculations of $\frac{1-E(e^{-\alpha S^{(r)}_{\tau+}})}{(1-e^{-\alpha})E(S^{(r)}_{\tau+})}$						
α	1	0.1	0.01	0.001	0.0001	0.00001
Martingale Method	0.5571	0.9202	0.9914	0.9991	0.9999	1.0
Equation (1.5)	0.5574	0.9233	1.0232	1.3682	14.29	641.1
Theorem 1	0.5571	0.9202	0.9915	0.9992	1.0	1.0

Table 2. The parameters for this table are $\lambda_i = p_i s$ where $p_1 = 0.6157$, $p_2 = 0.0269$, $p_3 = 0.3304$ and $p_4 = 0.0269$; $E[(S^{(r)}_{\tau+})^2]/2E[S^{(r)}_{\tau+}]$ and $E[S^{(r)}_{\tau+}]$ are calculated as a check for $s \downarrow 0$.

Random Walk	Mixture of Poisson		$s = 0.001$	$s = 0.01$	$s = 0.1$	$s = 1$
$\frac{E[(S^{(r)}_{\tau+})^2]}{2E[S^{(r)}_{\tau+}]} = 1.3632$	$\frac{E(S^2_{\tau+})}{2E(S_{\tau+})}$	Theorem 2	1.3642	1.3726	1.4567	2.2785
$E[S^{(r)}_{\tau+}] = 2.5387$	$E(S_{\tau+})$	Theorem 3	2.5395	2.5468	2.6210	3.4338

Let $Y_i = \exp(\theta S^{(r)}_i)$, where θ is a complex root of the equation $\phi(\theta) = \log E(\exp(\theta S^{(r)}_1)) = 0$. If $p_1 = 0.6157$, $p_2 = 0.0269$, $p_3 = 0.3304$ and $p_4 = 0.0269$, we have $\theta = -0.6787 + i(1.0351)$. It can be observed that $\{Y_i, i > 0\}$ is a Martingale and that $\tau_+ \equiv \inf\{n; S^{(r)}_n > 0\} = \inf\{n; Y_n > \exp(\theta)\}$ is a stopping time. The possible values for $S^{(r)}_{\tau+}$ are $\{1, 2, 3\}$. We apply $E(Y_{\tau_+}) = E(Y_1) = 1$ to solve $\pi_i = P(S^{(r)}_{\tau+} = i)$, and then $[1-E(e^{-\alpha S^{(r)}}_{\tau+})]/[(1-e^{-\alpha})E(S^{(r)}_{\tau+})]$, $E(S^{(r)}_{\tau+})$ and $E[(S^{(r)}_{\tau+})^2]$ can be calculated exactly. For this random walk example, we compare (1.5) in Tu and Siegmund (1999) with (2.1) for various α in Table 1. (1.5) breaks down for small α, while (2.1) improves substantially on it.

In Table 2, the Theorem 2 and Theorem 3 calculations for $E(S^2_{\tau_+})/2E(S_{\tau_+})$ and $E(S_{\tau_+})$ for the Poisson variables with $\lambda_i = sp_i$ are shown. In Table 3, various bounds on τ_+ are set in simulating $E(S^2_{\tau_+})/2E(S_{\tau_+})$ to show the efficiency that the theoretical results can provide.

3.3. Example 3: A sequential test example

Consider $\{x_i, i \geq 1\}$ to be i.i.d. with

$$x_1 = a_1 y_1 + a_2 y_2 - b_1 y_3 - b_2 y_4,$$

where y_i are independent Poisson variables with mean λ_i. The mean $\mu = E(x_i) > 0$ is chosen in such a way that $a_1 = b_1 = 1$, $a_2 = b_2 = Ma_1$, $\lambda_1 = 1.1$, $\lambda_3 = 1$,

Table 3. The Poisson parameters for this table are those of Table 2. In this simulation, various upper bounds (K) on τ_+ are set to show the values that τ_+ acquires to get a reasonable result.

	$\frac{E[S^2_{\tau_+\wedge K}]}{2E[S_{\tau_+\wedge K}]}$			
K	$s = 0.001$	$s = 0.01$	$s = 0.1$	$s = 1$
∞	1.378	1.370	1.457	2.257
20	2.014	2.028	1.870	2.241
100	2.056	1.979	1.488	2.262
200	1.858	1.846	1.456	2.366
400	1.726	1.582	1.460	2.266
600	1.644	1.460	1.410	2.283
800	1.629	1.457	1.441	2.291
1000	1.514	1.413	1.467	2.272
1500	1.444	1.369	1.509	2.230
2000	1.383	1.384	1.460	2.176

$\lambda_2 = \lambda_1/M$, and $\lambda_4 = \lambda_3/M$, where M is a positive integer referring to the ratio between high level and low level. These parameters are designed so that the four terms in the model have roughly equal means. The accumulated sum is $S_n = \sum_{i=1}^n x_i$. A sequential test stops when S_n reaches 15.

In this example, $T = \inf\{n : S_n \geq 15\}$. Theorem 2 can be applied to estimate $E(T) = E(S_T)/\mu$ by approximating the overshoot part $[E(S_T) - b]$ as $[E(S^2_{\tau_+})/2E(S_{\tau_+}) - 1/2]$. Theorem 3 can estimate the number of new-high points during the course as $E(S_T)/E(S_{\tau_+})$. The estimators are summarized in Table 4. In Table 5, the comparisons between the estimators and the simulations are shown for the waiting time, new high points and the overshoot for various M.

4. Discussion

The major challenge in calculating the overshoots is to manage the divergent points of the integrated function such that the orders of limit and integral are interchangeable. Complex analysis is applied to find the functions to solve these problems. The Theorem 1 formula for $[1-E(\exp(-\alpha S_{\tau_+}))]/[(1-\exp(-\alpha h))E(S_{\tau_+})]$ is more robust (over α) than that provided by Tu and Siegmund (1999). Our modification also makes it possible to reproduce the non-arithmetic result (Woodroofe (1979)) by letting h go to 0.

A concern in the three theorems is that the imaginary part of the log function is not well defined. For example, with $f(t)e^{i2\pi}$, $\log(f(t)e^{i2\pi}) = \log(f(t)) + i(2\pi)$, which means that the log function is ambiguous at multiples of $i(2\pi)$. Fortunately, this does not cause problems. The ambiguity of the log functions fall on the imaginary part, and the functions containing the imaginary parts of the log

Table 4. The estimators for overshoot, waiting time and the number of 'new-high' points.

	Overshoot	Waiting Time	New High Points
Estimator	$\frac{E(S^2_{\tau_+})}{2E(S_{\tau_+})} - 0.5$	$[b + \frac{E(S^2_{\tau_+})}{2E(S_{\tau_+})} - 0.5]/\mu$	$[b + \frac{E(S^2_{\tau_+})}{2E(S_{\tau_+})} - 0.5]/E(S_{\tau_+})$

Table 5. The performance of the estimators in Table 4.

	Overshoot		Waiting Time		New High Points	
M	Estim.	Simul.	Estim.	Simul.	Estim.	Simul.
5	2.0138	1.9858	85.0692	84.7034	5.432	5.4321
6	2.3317	2.3101	86.6585	86.7464	5.1281	5.1087
7	2.6579	2.6305	88.2897	88.0976	4.8895	4.8693
8	2.9917	2.9439	89.9586	89.6466	4.6991	4.6780
9	3.3323	3.3557	91.6614	90.8089	4.5448	4.5456
10	3.6790	3.7345	93.3949	92.9253	4.4185	4.4152

functions are all periodic odd functions, which means that the ambiguous part contributes 0 after integration.

The model in Example 3 of Section 3 can be applied to modeling accumulated gain or loss in an investment course or an insurance program. To make the applications of these results to more practical and interesting problems, the assumptions on the random variables x_i need to be more flexible. For example, the i.i.d. assumption could be relaxed to one of equal mean and equal variance, or to Markov random variables. The problems become more challenging, but the reward grows.

Acknowledgement

The author is very grateful to the referees and the associate editor for their thoughtful suggestions. The author would also like to thank Yuan-Fu Huang for providing help on programming.

References

Chan, H. and Zhang, N. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc.* **102**, 595-602.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

Spitzer, F. (1960). A Tauberian theorem and its probability interpretation. *Amer. Math. Soc.* **94**, 150-169.

Storey, J. D. and Siegmund, D. (2001). Approximate p-values for local sequence alignments: numerical studies. *J. Comput. Biology* **8**, 549-556.

Tu, I. and Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis, *Adv. Appl. Probab.* **31**, 510-531.

Woodroofe, M. (1979). Repeated likelihood ratio tests. *Biometrika* **66**, 454-463.

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC.

E-mail: iping@stat.sinica.edu.tw

(Received October 2006; accepted July 2007)