APPROXIMATING THE DISTRIBUTION OF THE MAXIMUM PARTIAL SUM OF NORMAL DEVIATES

Denis Conniffe*
and
John E. Spencer**

**The Economic and Social Research Institute, Dublin*

***Department of Economics, The Queen's University of Belfast, BT7 1NN, Northern Ireland*

January 1999

Working Paper No. 102

Working Papers are not for publication and should not be quoted without prior permission from the author(s).

Approximating the distribution of the maximum partial sum of normal deviates

By DENIS. CONNIFFE

The Economic and Social Research Institute, Dublin, Republic of Ireland

and JOHN. E. SPENCER

Department of Economics, The Queen's University of Belfast, BT71NN, Northern Ireland

SUMMARY

The largest partial sum of deviations from the mean is a statistic of importance in several areas of application, including hydrology and in testing for a change-point. Approximations to its distribution for the simple normal case have appeared in the literature, based either on functionals of Brownian motion asymptotics or on a methodology developed for boundary crossing problems in sequential analysis. The former approximation is inaccurate except for very large samples, while the latter is based on rather difficult theory. In this paper, we first review some early findings about exact moments and extend them somewhat. We then use these moments to fit simple Chi-squared and Beta approximations and show that they work very well.

Some key words: Change-point; Chi-squared and Beta approximations; Cusum; Hydrology

Address for correspondence:
The Economic and Social Research Institute, 4 Burlington Road, Dublin 4.

1. Introduction

The scaled largest partial sum of deviations from the mean

$$\text{M/s} = \frac{1}{s}\left[\max_k \left\{ \sum_1^k (x_i - \bar{x}) \right\} \right], \qquad (1)$$

where $1 \le k \le n$, $\bar{x} = \Sigma x / n$ *and* $s^2 = \Sigma(x - \bar{x})^2 / n$, with the summations taken 1 to n, is a statistic of importance in several areas of application. It seems to have first appeared in hydrology, in studies of reservoir storage (Hurst, 1951). The x's were variable inflows in successive time periods (often annual), with $\bar{x}$ a constant outflow, and it was called the "adjusted surplus". Along with the scaled minimum partial sum (the "adjusted deficit") and their difference (the "adjusted range", or R/s statistic) it played an important part in reservoir design and this stimulated statistical research on distributional properties, commencing with the asymptotic approach of Feller (1951). Substantial findings on the finite sample distribution followed, commencing with Anis & Lloyd (1953), which have been reviewed by Lloyd (1981).

The statistic (1) and several related statistics are also employed in testing for a change-point, that is, of testing the hypothesis that the x's are independently identically distributed against the alternative that at some r $(1 \le r \le n)$ the distribution changes. For the case of normal x's and an alternative of a single change in the mean, findings and discussion in James, James & Sigmund (1987) suggest that (1) is preferable when the change-point is towards the centre of the data set, rather than close to either end. They provided approximations to the null distributions of both M/s and the somewhat simpler statistic obtained by replacing s by a known standard deviation σ. These statistics can be given score type interpretations and have also been applied to other than the simple normal case. Pettitt (1980) considered zero-one data, while Ploberger & Kramer (1992) use (1) to test for parameter constancy by replacing the $x_i - \bar{x}$ by regression residuals, giving a cusum statistic like that of Brown, Durbin & Evans (1975), but based on ordinary rather than recursive residuals.

Statistics similar to (1) also arise in boundary crossing problems in sequential analysis. Indeed, it was the methodology developed in that field (Siegmund, 1985, 1986) that was utilised to derive the approximate distributions of M/σ and M/s in James et al (1987). See also James, James and Siegmund (1988). The methodology is quite difficult, involving large deviation theory and discrete time

corrections to functionals of Brownian motion type approximations to boundary crossing probabilities, although the final formula for a significance probability is simple for M/σ and tractable, given numerical computation, for M/s. The approximating distribution employed by Ploberger & Kramer (1992) uses the familiar Brownian bridge approximation to partial sums of deviations to derive critical points. However, unless n is very large the tail probabilities can be very inaccurate.

In many fields of statistics the most widely-used, and often the soundest (see, for example, Cox & Hinkley (1974), pp. 462-465) approach to approximating a complicated distribution is to fit a simpler distribution, that has much the same range and general shape, by equating moments. In this paper, we first review the existing results on the exact moments of M/σ and M/s and extend them somewhat. We then show that for the simple normal case Chi-squared and Beta approximations are easily fitted. Through a Monte Carlo simulation study, we confirm these approximations are very accurate and compare well with the James et al formulae and are much superior to the Ploberger & Kramer critical values. We conclude with some comments on possible extensions to econometric models.

2. Exact moments

Anis & Lloyd (1953) investigated the distribution of

$$\max_k\left(\sum_1^k x_i\right),$$

with $1 \le k \le n$ and the x's independent standard normal and found its exact expectation. Anis (1955) found the exact variance and continued (Anis, 1956) to obtain a recurrence relation for higher moments. Solari & Anis (1957) considered the distribution of M/σ

$$\frac{1}{\sigma}\left[\max_k\left\{\sum_1^k (x_i-\bar{x})\right\}\right]$$

where the x's are again independent with arbitrary mean, but known variance. They showed the distribution has a 'spike' at the origin, the probability of all the partial sums being negative or zero equalling n^{-1}, and they obtained the first two moments. The expectation (as later slightly rearranged by Boes & Salas-La Cruz (1973)), is

$$\sqrt{\frac{1}{2n\pi}}\sum_{j=1}^{j=n-1}\sqrt{\frac{n-j}{j}} \qquad (2)$$

and the second moment is

$$\frac{1}{6}\left\{\frac{n^2-1}{n}+\frac{\sqrt{n}}{2\pi}\sum_{i=2}^{n-1}\sum_{j=1}^{i-1}\frac{i(2i-n)}{[(n-i)j^3(i-j)^3]^{1/2}}\right\}. \qquad (3)$$

These have been tabulated by Solari & Anis for a range of n, but are obviously easily computed.

The expectation of (1), M/s, could now have easily been deduced from (2) if Solari & Anis had employed a result of Geary (1933). Geary showed that, for normal samples, ratios with s^2 as the denominator are independent of s^2 if they are homogeneous of degree zero. Hence, in such cases, moments of ratios are obtained by dividing the moments of the numerators by those of the denominators. The statistic (1) is homogeneous of degree zero and so, given normality, is independent of its denominator. So the expectation of the numerator, which is (2) by σ, is the product of the expectation of the ratio and the denominator. As is well known (for example, exercise 17.6 of Kendall & Stuart, 1967, vol. 2, p.32)

$$\frac{1}{\sigma}\mathrm{E(s)}=\frac{\sqrt{2}}{\sqrt{n}}\,\frac{\Gamma\,(n/2)}{\Gamma\{(n-1)/2\}} \qquad (4)$$

and dividing (2) by (4) immediately gives

$$\frac{1}{2}\frac{1}{\sqrt{\pi}}\,\frac{\Gamma\,\{(n-1)/2\}}{\Gamma(n/2)}\sum_{j=1}^{j=n-1}\sqrt{\frac{n-j}{j}}\ . \qquad (5)$$

Instead, the exact mean of M/s remained unknown until Anis and Lloyd (1976), using a theorem of Spitzer (1956) showed, with substantial manipulation, that it is (5). Using the same approach, the second moment of M/s, which has not previously appeared in the literature, is (3) divided by the expectation of s^2/σ^2, which is (n-1)/n. Rearranging to a more convenient computational form, it is

$$\frac{1}{6}\left[n+1+\frac{n\sqrt{n}}{(n-1)\pi}\sum_{i=2}^{n-1}\left\{\frac{2i-n}{\sqrt{n-i}}\sum_{j=1}^{i-1}\frac{1}{j^{3/2}\sqrt{i-j}}\right\}\right]. \qquad (6)$$

3. Approximating by equating moments

Since the distributions of M/σ and M/s have a spike at the origin, what will be approximated are the distributions conditional on non-zero statistics. The means and second moments are then n/(n-1) multiplied by (3) and (4) for M/σ and n/(n-1) by (5) and (6) for M/s. The ascertained probability that a critical point is exceeded in an approximating distribution can then be multiplied by (n-1)/n to allow for the probability mass at the origin.

There is an important difference between the M/σ and M/s statistics in that the latter is bounded and functions of bounded variables are often approximated by Beta distributions. From Mandelbrot (1972), it is clear that $M / s \leq n/2$, and so

$$\frac{4}{n^2}\frac{M^2}{s^2} = \frac{4}{n}\frac{M^2}{\sum(x_i - \bar{x})^2} \tag{7}$$

appears to have the right dimensions for a Beta, with a range from zero to one, the denominator a sum of squares of deviations and the Beta ratio property that the ratio is independent of the denominator.

For known σ, which may be taken as unity, the corresponding approximation relates $4M^2/n$ to a chi-squared. A simple one moment approximation would then take $4M^2/n$ as chi-squared with (non-integer) degrees of freedom estimated as either f = (3) multiplied by 4/(n-1), or f * obtained by equating (2) multiplied by $2\sqrt{n}/(n-1)$ to the 'half' moment of chi-squared

$$\sqrt{2}\frac{\Gamma(f/2+.5)}{\Gamma(f/2)}.$$

The comparison of the two estimates in Table 1 for a selection of values of n shows little difference and provides reassurance about the plausibility of the chi-squared approximation. In the next section we will use f, rather than f*, because it is simplest.

Table 1. *'Degrees of freedom' f and f**

Sample size	10	20	30	40	60	80	100
f	1.34	1.49	1.56	1.61	1.67	1.71	1.74
f*	1.36	1.50	1.57	1.61	1.67	1.71	1.74

A two moment approximation to $a\chi_f^2$ or to χ_f^{2C} would be possible, of course, but as Table 1 suggests and the simulations in the next section will confirm, it is hardly worthwhile. It may be worth noting that Solari & Anis (1957) showed that (3), $\rightarrow n/2 - \sqrt{n}$ as $n \rightarrow \infty$, so $f \rightarrow 2$ as $n \rightarrow \infty$, although it is well short of this for n = 100 and converging only slowly.

Returning to the case of unknown σ, let c_1 denote (5) multiplied by 2/(n-1) and c_2 denote (6) multiplied by $4/(n^2-n)$. A one moment approximation to a Beta, with parameters p and q, would take p + q = (n-1)/2 and, via $c_2 = p/(p+q)$, $p = c_2(n-1)/2$. A two moment approximation uses $c_2 = p/(p+q)$ and solves for p from

$$c_1 = \frac{\Gamma(p/c_2)\Gamma(p+.5)}{\Gamma(p/c_2+.5)\Gamma(p)}.$$

4. Comparing approximations

The asymptotic (via the Brownian Bridge) approximation to the probability, assuming $\sigma = 1$, that

$$\frac{1}{\sqrt{n}}\left[\max_k\left\{\sum_1^k (x_i - \bar{x})\right\}\right] > b, \tag{8}$$

used by Ploberger and Kramer (1992) is

$$\sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2b^2}. \tag{9}$$

The large deviation approximation given by James, James and Siegmund (1987) is

$$e^{-2\left(b+\frac{.583}{\sqrt{n}}\right)^2}. \tag{10}$$

Of course, the probability of (8) is the same as that of

$$\frac{4}{n}\left[\max_k\left\{\sum_1^k (x_i - \bar{x})\right\}\right]^2$$

exceeding $4b^2$, which is approximated by chi-squared. For large b and $n \to \infty$ all three become $\exp(-2b^2)$, since the large deviation probability for a chi-squared exceeding $4b^2$ is (for example, Abramowitz & Stegun, 1972, p. 941)

$$\frac{1}{\Gamma(f/2)}(2b^2)^{f/2-1}e^{-2b^2}.$$

However, in finite samples it turns out that the probability given by chi-squared is close to the true probability of (8) and to (10), but quite different from (9). In Table 2, the first column in each sample size sub-table contains the χ_f^2 'critical values' for $n\alpha/(n-1)$ with $\alpha = .5, .4, .3, .2, .1, .05, .025, .01$. The second column, labelled α^* gives the actual proportions of times these values were exceeded in a Monte Carlo simulation with 6000 replications for each sample size and, obviously, the closer α^* to α, the better the chi-squared approximation. The distribution of

$$-\frac{1}{\sqrt{n}}\left[\min_k\left\{\sum_1^k (x_i - \bar{x})\right\}\right]$$

is identical to that of $M/\sqrt{n}$, so that proportions were also estimated from this statistic. This more than doubled the effective simulation replication, because the maximum partial sum and minus the minimum partial sum are negatively correlated and so constitute antithetic variates. The third column contains the value given by (10) with b replaced by half the square root of the critical chi-squared

value and is labelled JJS. The fourth column contains the corresponding values given by (9) and is labelled PK. The results presented are for the sample sizes n =10, 20, 30, 40, 60, 80 and 100. A larger set were actually employed in the study, but the results for 50, 70, etc. conform fully with those presented.

Table 2. *Performance of the chi-squared approximation – the known variance case*

Sample size		n = 10				n = 20				n =30			
α	χ_f^2	α^*	*JJS*	*PK*	χ_f^2	α^*	*JJS*	*PK*	χ_f^2	α^*	*JJS*	*PK*
.5	.61	.512	.515	.499	.82	.508	.507	.493	.91	.503	.505	.488
.4	.93	.411	.410	.486	1.17	.403	.405	.466	1.28	.405	.404	.453
.3	1.37	.307	.305	.441	1.65	.302	.303	.402	1.78	.299	.302	.383
.2	2.03	.198	.200	.345	2.35	.196	.200	.300	2.50	.199	.200	.280
.1	3.22	.093	.096	.198	3.59	.096	.098	.166	3.76	.095	.099	.152
.05	4.45	.047	.046	.108	4.86	.050	.048	.088	5.05	.047	.048	.080
.025	5.71	.023	.022	.058	6.15	.025	.023	.046	6.36	.024	.024	.042
.01	7.40	.008	.008	.025	7.88	.011	.009	.019	8.11	.010	.009	.017

Sample size		n = 40				n = 60				n =80			
α	χ_f^2	α^*	*JJS*	*PK*	χ_f^2	α^*	*JJS*	*PK*	χ_f^2	α^*	*JJS*	*PK*
.5	.97	.502	.504	.484	1.04	.493	.503	.478	1.09	.503	.502	.474
.4	1.35	.398	.403	.444	1.43	.398	.403	.433	1.49	.405	.402	.426
.3	1.86	.299	.302	.371	1.95	.301	.302	.357	2.01	.302	.302	.348
.2	2.59	.202	.201	.269	2.69	.202	.201	.256	2.76	.197	.201	.247
.1	3.87	.099	.099	.144	3.99	.097	.099	.136	4.07	.100	.099	.130
.05	5.17	.049	.049	.075	5.31	.049	.049	.070	5.40	.050	.049	.067
.025	6.48	.024	.024	.039	6.64	.024	.024	.036	6.73	.025	.024	.034
.01	8.24	.010	.009	.016	8.40	.009	.010	.015	8.51	.010	.010	.014

Sample size n=100

α	χ_f^2	α^*	JJS	PK
.5	1.12	.503	.503	.472
.4	1.52	.403	.402	.421
.3	2.05	.301	.301	.342
.2	2.81	.203	.201	.242
.1	4.12	.102	.100	.127
.05	5.45	.052	.050	.065
.025	6.79	.027	.024	.033
.01	8.58	.010	.010	.014

The chi-squared approximation generally fits well over all 'critical values', so that it can be used to evaluate P-values as well as test hypotheses, but the tail points for α =.1, .05, .025 and .01 are of most interest. The corresponding α^* values are obviously close to these α, but the matter deserves a little

more examination. The James et al (1987) probability formula (10) applied to the chi-squared points give very similar values to the simulation results. If the JJS tail values were consistently substantially closer to the α^* than the latter were to the α, the chi-squared approximation would be inferior to use of (10). However, except for n=10, this is not consistently so and even then the difference is slight. It may be worth remarking that for tail values and n as low as 10, accurate probabilities can be obtained from Bonferroni bounds in the manner of Worsley (1982). All this is not to claim that the chi-squared approximation is superior to (10), but it is of comparable performance.

The Ploberger & Kramer (1992) formula (9) gives higher values in the tails – very much so for low n, but still appreciable at n=100. So using critical values based on (9) would imply a true size of test well below the nominal α. Ploberger & Kramer reported a Monte Carlo simulation for n=120, where the proportion exceeding a nominal 5% (two tail) point was actually .0378. The situation is much worse at lower n, however. In our simulation, for example, the estimated probability of (8) with b = 1.36, which is the nominal 2.5% (one tail) point based on (9), was .01 for n=30 and .006 for n=20.

Turning to the case of σ unknown, the James et al formula for the probability that

$$\frac{1}{\sqrt{n}\,s}\left[\max_k\left\{\sum_1^k (x_i - \bar{x})\right\}\right] > b$$

is

$$\left\{1-4\frac{b^2}{n}\right\}^{(n-3)/2} V\left\{\frac{4b/\sqrt{n}}{\sqrt{(1-4b^2/n)}}\right\}, \tag{11}$$

where

$$V(z) = \frac{2}{z^2}\exp\left\{-\sum_1^\infty \frac{1}{i}\Phi(-\frac{1}{2}z\sqrt{i})\right\}, \tag{12}$$

with Φ the standard normal distribution function.

The two Beta based approximations were described in the last section. The simple single moment approximation will be denoted B1 and the two moment approximation B2. The 'critical' tail values for both were obtained from the inverse of the Beta integral (now a standard feature in most statistical computing packages) and compared in a simulation study, which was again based on 6000 replications and used both the maximum and minus the minimum partial sums.. For consistency, the corresponding 'critical' JJS values were obtained from solution of (11) for b, given tail probabilities. As this is not a trivial computation, the approximation, used in James et al (1987), of $V(z) \approx \exp(-.583z)$ and valid

for $0 \leq z \leq 2$ was employed in all cases except for n =10, with α =.025 and .01, when z exceeded 2. The Ploberger & Kramer critical values were also investigated in the study, but as in the case of known σ, proved so inaccurate that they will not be further reported on.

The three sets of 'critical' values are generally very close together. For lower sample sizes, the pattern is that the B1 value is slightly higher (further to the right) than the B2 value, which in turn is very slightly higher than the JJS value, but as sample size increases the B1 value falls very slightly below the B2 value. The important issue is how they compare in terms of 'true' probabilities as estimated in the Monte Carlo study. The results are shown in Table 3.

Table 3. *Performance of the Beta approximations – the unknown variance case*

	α =.1			α =.05			α =.025			α =.01		
n	B1	B2	JJS	B1	B2	JJS	B1	B2	JJS	B1	B2	JJS
10	.096	.100	.105	.046	.050	.055	.020	.023	.023	.008	.010	.011
20	.093	.094	.094	.044	.046	.046	.019	.020	.021	.006	.007	.008
30	.096	.097	.097	.047	.048	.048	.024	.024	.025	.009	.009	.010
40	.098	.099	.100	.047	.048	.048	.020	.020	.020	.007	.007	.008
60	.099	.098	.100	.050	.050	.051	.025	.024	.025	.009	.009	.009
80	.097	.097	.097	.051	.050	.051	.026	.026	.026	.010	.010	.011
100	.099	.099	.099	.051	.051	.051	.025	.025	.025	.010	.010	.011

The distributions of 6000 maxima and minima were generated independently for the different sample sizes, but of course are kept the same for the critical values given sample size and so, for example, the consistent slight shortfall from nominal levels for n =20 contains a sampling effect. Overall, there is nothing to choose between B2 and JJS as regards accuracy of test size. B1 is just as accurate as B2 for n > 40, but even for n =20 the very slight underestimation relative to B2 could hardly be considered of practical importance. The underestimation may be more appreciable at n =10, though still small, but the earlier comment about Bonferroni inequalities applies to the unknown σ case too.

Just as for chi-squared in the case of known σ, the simulation showed that the Beta approximations were accurate in giving probabilities well outside the tail, so that computation of P values is quite feasible .

5. Concluding remarks

The James et al (1987) approximations to the distributions of M and of M/s are very good and the chi-squared and Beta approximations obtained in this paper cannot really improve on them. But they are as good and are derived by the familiar method of fitting curves by moments using longstanding results from statistical hydrology and a simple application of Geary (1933). But besides showing the

relevance of these rather neglected results, the approach of this paper may be relevant in fields where testing for parameter stability is important. For example, in econometrics the topic of structural change and the need to test for it has received and continues to receive much attention.

The authors providing the methodology for such investigations (Ploberger and Kramer, 1992; Andrews, 1993; among others) rely heavily on approximations based on functionals of Brownian motion in spite of the associated inaccuracy. This may be because the theory underlying the James et al approximations is considered difficult and the task of applying it the maximum of a partial sum of residuals (or some related statistic) from a complex econometric model, instead of from a simple normal distribution, may be seen as daunting. Perhaps the possibility of obtaining accurate approximations through traditional fitting by moments may be considered less so. The task is still demanding, of course, as at least one moment of the statistic must be determined, or at least approximated.

REFERENCES

Abramowitz , M. & Stegun, I. A. (1972). *Handbook of mathematical functions*. New York: Dover.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**, 821-56.

Anis A. A. & Lloyd E. H. (1953). On the range of partial sums of a finite number of independent normal variables. *Biometrika* **40**, 35-42.

Anis A. A. (1955). The variance of the maximum of partial sums of a finite number of independent normal variates. *Biometrika* **42**, 96-101.

Anis A. A. (1956). On the moments of the maximum of partial sums of a finite number of independent normal variables. *Biometrika* **43**, 79-84.

Anis A. A. & Lloyd E. H. (1976). The expected value of the adjusted rescaled Hurst range of independent normal summands. *Biometrika* **63**, 111-6.

Boes, D. C. & Salas-La Cruz, J. D. (1973). On the expected range and expected adjusted range of partial sums of exchangeable random variables. *J. Appl. Prob.* **10**, 671-7.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time (with discussion). *J. R. Statist. Soc. B* **37**, 149-92

Cox, D. R. & Hinkley D. V. (1974). *Theoretical Statistics,* London: Chapman and Hall.

Feller, W. (1951). "The asymptotic distribution of the range of sums of independent random variables", *Ann. Math. Statist.* **22**, 427-32.

Geary, R. C. (1933). A general expression for the moments of certain symmetrical functions of normal samples. *Biometrika* **25**, 184-6.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Am. Soc. Eng.* **116,** 770-99.

James, B., James K. L., & Siegmund, D. (1987). Tests for a change-point. *Biometrika* **74,** 71-83.

James, B., James K. L., & Siegmund, D. (1988). Conditional boundary crossing probabilities, with applications to change-point problems. *Ann. Prob.* **16,** 825-39.

Kendall, M. G. & Stuart, A. (1967). *The Advanced Theory of Statistics* vol. 2. London: Griffin.

Lloyd E. (1981). Stochastic hydrology: an introduction to wet statistics for dry statisticians. *Commun. Statist. Theor. Meth. A* **10(15),** 1505-22.

Mandelbrot, B. (1972). Statistical methodology for non periodic cycles: from the covariance to R/S analysis. *Ann. Econ. Soc. Meas.* **1,** 259-90.

Pettitt, A. N. (1980). A simple cumulative sum type statistic for the change-point problem with zero-one observations. Biometrika **67**, 79-84.

Ploberger, W. & Kramer, W.(1992). The cusum test with ols residuals. *Econometrica* **60**, 271-85.

Siegmund, D. (1985). *Sequential Analysis: tests and confidence intervals*, Springer-Verlag: New York.
Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *Ann. Statist.* **14**, 361-404.
Solari, M. E. & Anis, A. A. (1957). The mean and variance of the maximum of the adjusted partial sums of a finite number of independent normal variates. *Ann. Math. Statist.* **28**, 706-16.
Spitzer, F. (1956). A combinatorial lemma and its application to probability theory. *Trans. Am. Math. Soc.* **82**, 323-39.
Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69**, 297-302.