The Economic and Social Review, Vol. 31, No. 3, July, 2000, pp. 237-248

Approximating the Distribution of the R/s Statistic[*]

DENIS CONNIFFE
The Economic and Social Research Institute

and

JOHN E. SPENCER
The Queen's University of Belfast

Abstract: The R/s statistic, used for many years in hydrology, is increasingly employed in economics, although deficiencies in knowledge about its exact distribution have inhibited progress. Harrison and Treacy (1997) described some applications where R/s arises as a test statistic and they derived close to exact critical values for conventional (5 per cent etc.) significance levels for a range of sample values through Monte Carlo simulation. This paper examines two approaches. One is a simple adjustment to the asymptotic distribution that improves its upper tail accuracy greatly and the other is an approximation to the whole distribution, easily computed and suitable for "P-value" calculation, which is also reasonably precise in the upper tail. The Harrison and Treacy values and Monte Carlo simulation are used to confirm accuracy.

I INTRODUCTION

For a sample of size n the R/s statistic is

$$\mathrm{R/s} = \frac{1}{s}\left[\max_k\left\{\sum_1^k (x_i - \bar{x})\right\} - \min_k\left\{\sum_1^k (x_i - \bar{x})\right\}\right], \tag{1}$$

where $1 \le k \le n$, $\bar{x} = \sum x / n$ and $s^2 = \sum (x - \bar{x})^2 / n$, with the summations taken 1 to n. Since the sum of deviations over the whole sample is zero, the maximum

[*]The authors are grateful for some financial support from the Social Science Research Council. We also thank the anonymous referees for helpful comments.

partial sum in (1) is either positive or zero and the minimum partial sum is either negative or zero. So (1) is always positive. It was originally employed (Hurst, 1951) in studies of reservoir storage, where the x's were variable inflows in successive time periods (often annual) and $\bar{x}$ was the constant outflow. Naturally, much study took the x's as normally and independently distributed, at least as a null hypothesis, but many complications have been introduced. Developments in the hydrological literature are described by Lloyd (1981).

The statistic (1) and others similar to it (for example, the maximum partial sum divided by the sample standard deviation) are also employed in testing for change points and in boundary crossing problems in sequential analysis. The statistic was suggested for testing for long-term dependence in economic and financial time series by Mandelbrot (1971, 1972), but only in recent years has it begun to appear frequently in the economics literature. Harrison and Treacy (1997) provide an account of applications and of why the R/s statistic seems appropriate and Harrison and Treacy (1998) discuss its use in testing for parameter instability.

It is intuitively evident from (1) that the exact, small sample, distribution of the R/s statistic is complicated. For the case of independent x's (not necessarily normally distributed) Feller (1951) found the asymptotic distribution of R/s divided by the square root of n. Using a more familiar modern terminology and notation than he did, the asymptotic distribution is that of the range of a Brownian Bridge and the distribution function is

$$P\left(\frac{1}{\sqrt{n}}\frac{R}{s} \le V\right) = 1 - 2\sum_{j=1}^{\infty}(4j^2V^2 - 1)e^{-2j^2V^2}. \tag{2}$$

From early on, it was appreciated that the asymptotic distribution was inaccurate in small samples. For example, the mean of the asymptotic distribution is 1.253 (so that means of repeated calculations of R/s with the same sample size n should be $1.253\sqrt{n}$), but observations in hydrology did not seem to support means of R/s being proportional to $\sqrt{n}$. This contradiction is called "the Hurst effect" and one possible explanation (see, for example, Anis and Lloyd (1975)) is that the true (or finite sample) expectation differs substantially from the asymptotic mean.

For finite n and assuming the x's independently drawn from the same normal distribution with known variance σ^2, Solari and Anis (1957) showed that the expectation of the numerator of (1) is

$$\sigma\sqrt{\frac{2}{n\pi}}\sum_{j=1}^{n-1}\sqrt{\frac{n-j}{j}}. \tag{3}$$

The expectation of (1) could easily have been deduced from this result if the authors had appreciated the significance of a result by Geary (1933). Geary showed that, for a normal sample, ratios with s^2 as the denominator are independent of s^2 if they are homogeneous of degree zero (in polar co-ordinates, if they are functions of the polar angles only). Then he obtained moments of ratios by dividing the moments of the numerators by those of the denominators. The R/s statistic (1) is homogeneous of degree zero and so, given normality, is independent of its denominator. So the expectation of the numerator, which is (3), is the product of the expectation of the ratio and the denominator. As is well known (for example, exercise 17.6 of Kendall and Stuart, 1967, Vol. 2, p. 32)

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

and dividing (3) by (4) immediately gives

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

In fact, discovery of the exact mean of R/s had to wait until Anis and Lloyd (1976), who used a theorem of Spitzer (1956) and, with substantial manipulation, showed that it is (5). The "moments of ratios equalling ratios of moments" device is mentioned here, not only as a point of historical interest, but because it will be employed again in the paper.

Harrison and Treacy (1997) investigated the exact finite sample distribution of R/s by Monte Carlo simulation for a range of sample values, estimating moments and fitting a Beta approximation. They used this to estimate (very accurately) the critical tail points corresponding to the conventional (10 per cent, 5 per cent, etc.) significance levels. Their critical values were much lower than the asymptotic values for quite realistic (in economics, at least) sample sizes and an appreciable discrepancy persisted even up to large sample sizes of 500. So for tests where the null distribution is that of R/s with the x's a normal iid sample, the nominal significance levels of the asymptotic points can be very wrong. The probability of rejecting the null hypothesis when it is true may be far lower than the nominal levels indicate. A corollary is that the power of tests when the null is untrue may be poor.

The situation is illustrated in Table 1, which shows the tail "probabilities" produced by formula (2) when the Harrison and Treacy critical values (from their Table 10) for a selection of sample sizes are put equal to V.

Table 1: *Probability of >V according to Asymptotic Approximation*

True α	*n=20*	*n=40*	*n=60*	*n=80*	*n=100*
.10	.348	.248	.211	.192	.179
.05	.243	.153	.124	.110	.101
.01	.115	.052	.037	.030	.026

For n=20 formula (2) overestimates by a factor of 11 for the 1 per cent point and nearly 5 for the 5 per cent point. Even at n=100 the overestimation factors are still large at 2.5 and 2.

While the Harrison and Treacy critical values are very accurate for the conventional significance levels and sample sizes they considered, there are still inconveniences for the potential user. For other sample sizes, interpolation from their published tabulated points is one possibility, but may be thought inconvenient and gives scope for mistakes. Harrison and Treacy do mention they have more detailed tabulations than presented in the paper, which they are ready to make available. However, complete tabulation of critical point values for a range of significance levels and all possible sample sizes up to, say, 500 would require clumsy tables. More importantly perhaps, many researchers like to see a "P value" — the probability that the value of their test statistic would have been exceeded under the null. This could be estimated from the Harrison and Treacy work by calculating the mean, estimating the second, third and fourth moments for the relevant n via their Table 8 and proceeding to estimate the four parameters involved in their Beta approximation. Again, this could be thought inconvenient and relatively prone to error.

This paper obtains new approximations to the exact distribution. While these cannot improve on the virtually exact critical values found by Harrison and Treacy for the conventional significance levels at their sample sizes, they can easily be employed to obtain, at least approximately, "P values" for any sample size. Two approaches are employed. The first, which will be described in Section II, can be seen as an adjustment to the asymptotic distribution (2) that not only simplifies it, but improves its tail accuracy greatly. The second, presented in Section III, is a simple two moment Beta type approximation to the whole distribution, similar to the Conniffe and Spencer (2000) approach in the case of the maximum partial sum. Other approximations and approaches are also possible and are discussed briefly in Section IV as are desirable directions for further research.

II A LARGE DEVIATION TYPE APPROXIMATION

The distribution of a statistic z, based on a sample of finite size n of x's, may be complicated over most of its range, but simplify in its tail region, because various terms may become negligible for large z. The limiting form of this tail distribution as n becomes increasingly large then gives an "extreme value" or "large deviation" approximation. The familiar asymptotic approximations employed in econometrics (of the Central Limit Theorem type) first approximate the whole distribution and then go to the tail values for critical points. Generally, the two approaches are not equivalent and the large deviation method often gives much more accurate approximations to tail area probabilities. On the other hand, the approximations may be poor outside tail areas and the large deviation probability formula can depend on the distribution of the x's. Central Limit Theorem type approximations are usually more robust to distributional assumptions.

The R/s statistic involves the largest and smallest partial sums of deviations from the sample mean. The probability that extrema of partial sums cross boundaries is an important topic in several branches of statistics including quality control, renewal theory and sequential analysis and approximations have been based on both large deviations and on Brownian motion. See Siegmund (1985, 1986). In some cases, the large deviation approximation can appear as a straightforward (though dependent on sample size) adjustment to the tail points derived from Brownian motion arguments. The adjustment can be given various interpretations – as a correction for discreteness to the "continuous time" context of Brownian motion, or as an analogue of an Edgeworth adjustment to the Central Limit Theorem — as described in Siegmund (1985, Chapter 10). However, for the purposes of this paper, the important point is the suggestion that a fairly simple, but effective, correction to the asymptotic critical points may be possible.

A shortcut to a correction is suggested by considering the accurate large deviation approximation to the probability that

$$\frac{1}{\sqrt{n}}\frac{1}{s}\left[\max_k\left\{\sum_1^k(x_i-\bar{x})\right\}\right]>b, \tag{6}$$

obtained by James, James and Siegmund (1987), for normal iid x's. The approximation was

$$e^{-2\left(b+\frac{.583}{\sqrt{n}}\right)^2}. \tag{7}$$

Now the asymptotic (Brownian Process crossing a boundary) approximation to the probability of (6), used, for example, by Ploberger and Kramer (1992) is

$$\textstyle\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2b^2}. \tag{8}$$

Just as for the asymptotic distribution (2) for the R/s statistic, (8) gives a poor approximation to the distribution of (6). Ploberger and Kramer conducted a Monte Carlo simulation for n = 120 and found the true probability of exceeding a nominal 5 per cent point to be .0378. For small sample sizes the situation is far worse, as shown in Conniffe and Spencer (2000).

For a large value of b, (8) is just equal to the first term of the infinite series, which is

$$e^{-2b^2},$$

and "correction" of b by an amount $.583/\sqrt{n}$, gives the good approximation (7). Now, from (2)

$$P\left(\frac{1}{\sqrt{n}}\frac{R}{s} > V\right) = 2\sum_{j=1}^{\infty}(4j^2V^2 - 1)e^{-2j^2V^2}$$

which for large V is

$$2(4V^2 - 1)e^{-2V^2}. \tag{9}$$

and analogy with the approximation (7) to the probability of (6) suggests

$$2\left[4\left(V + \frac{c}{\sqrt{n}}\right)^2 - 1\right]e^{-2\left(V+\frac{c}{\sqrt{n}}\right)^2}, \tag{10}$$

with c a constant to be determined. An estimate of c can be obtained from any value in Table 1 in the following way. Equate (9) to the nominal α and solve for V. Then equate it to the true probability of exceedance and solve. The difference, d say, between the two solutions is an estimate of c divided by root n. There is some variation in the estimates obtained depending on choice of α and n, so a "pooled" estimate was obtained by "regression" of d on $1/\sqrt{n}$. This resulted in a value of about 1.4.

Now this argument has been heuristic and, without theoretical justification for (10) or for the estimate of c, it is not possible to make *a priori* assessments of the order of accuracy of the approximation. It may seem surprising that the single, fairly simple expression (10), with c = 1.4, can give a much better approximation to the exact tail distribution than (2) can, but it is so. Table 2 gives the values of (10) when the Harrison and Treacy critical values, for the same range of sample sizes as before, are substituted for V.

Table 2: *Probability of >V According to Large Deviation Approximation*

True α	*n=20*	*n=40*	*n=60*	*n=80*	*n=100*
.10	.087	.088	.089	.090	.090
.05	.052	.048	.047	.047	.047
.01	.019	.013	.011	.011	.010

Comparison of these values with those of Table 1 shows a dramatic improvement. The values are close enough to the true α (with the possible exception of the 1 per cent point with n=20) to justify the practical employment of (10) for 10 per cent, 5 per cent and 1 per cent significance tests. However, (10) could not be expected to give accurate "P Values" outside this tail area.

Note that in all of the development it has been assumed that what is of interest is

$$P\left(\frac{1}{\sqrt{n}}\frac{R}{s} > V\right)$$

for large V, that is, probabilities in the right hand tail of the distribution. The formulae would not apply to small V, that is, to the left hand side of the distribution. However, the hypothesis test situations in which the R/s statistic would typically be employed imply upper tail rejection regions (as is almost always the case with a test statistic, such as chi-squared, which must be positive).

III AN APPROXIMATION TO THE WHOLE DISTRIBUTION

One of the most widely-used, and often the soundest (see, for example, Cox and 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. This assumes some moments of the complicated distribution are known, or can at least be adequately approximated. The expectation of R/s is available, as given by (5). For the second moment, using Geary (1933),

$$E\left(\frac{R}{s}\right)^2 = \frac{E(R^2)}{E(s^2)} = \frac{\text{var}(R) + [E(R)]^2}{E(s^2)} \tag{11}$$

and E(R) is known exactly from (3), while the denominator is $(n\text{-}1)\sigma^2/n$. As regards the variance of R, various authors, from Solari and Anis (1957) to Harrison and Treacy (1997), have noted that the finite sample variance approaches Feller's (1951) asymptotic variance fairly quickly with increasing sample size, although the finite mean and second moments approach their corresponding asymptotic values only slowly. Feller's variance formula for R is .074 nσ^2. Based on Monte Carlo simulations, Phien, Arbhabhirama and Sutabutr (1979) suggested a correction to (.074 n +.062) σ^2. Although very slight unless samples are small, the correction is so simple as to merit inclusion. Inserting in (11) gives quite a good approximation to the second moment of R/s, as comparison with the simulation findings in Table 1 of Harrison and Treacy (1997) can verify.

As noted by Mandelbrot (1972), $1 \le R/s \le n/2$, and functions of bounded variables are often approximated by Beta distributions. So the variable

$$y = \frac{4}{n^2}\frac{R^2}{s^2} = \frac{4}{n}\frac{R^2}{\sum(x_i - \bar{x})^2}$$

appears to have the right dimensions for a Beta, with the denominator a sum of squares and the Beta ratio property that the ratio is independent of the denominator. The range is from 4/n^2, which can be taken as effectively zero, to unity. An exact half moment of the true distribution follows from (5) multiplied by 2/n (call it $c_{.5}$) and it can be equated to the half moment of a Beta (p, q) distribution

$$\frac{\Gamma(p+q)\Gamma(p+0.5)}{\Gamma(p+q+0.5)\Gamma(p)}.$$

An (approximate) first moment of y is (11), using the var(R) approximation, multiplied by 4/n^2 (call it c_1) and it can be equated to the first moment p/(p + q). Then, substituting for q in the half moment, the value of p is the solution of

$$\frac{\Gamma\left(\frac{p}{c_1}\right)\Gamma(p+0.5)}{\Gamma\left(\frac{p}{c_1}+0.5\right)\Gamma(p)} = c_{.5}, \tag{12}$$

giving a two moment approximation to a Beta. Nowadays, standard econometrics packages can evaluate Gamma functions and facilitate summing series. Otherwise solution of (12) would be tedious, as indeed would calculation of (5). Good packages (for example, Shazam (1997)) can also integrate the density functions of the commonly encountered probability distributions, including the Beta. For example, for n=40 the values of p and q turn out to be 4.266 and 29.886. If R/s is 9.69, then y is .2347 and the corresponding Beta cumulative distribution value is .95, so that there is just "significance" at 5 per cent. The immediate question is how accurate are "P Values" calculated through this Beta approximation. As Harrison and Treacy (1997) only provided α points for the tail area, this cannot be adequately assessed by the device, used in the previous section, of inserting their critical values into the probability formula. Instead, the Beta values corresponding to $\alpha = .5, .4, .3, .2, .1, .05$ and $.01$ were computed (by inverting the Beta integral) for the p's and q's corresponding to a range of sample sizes. Then a Monte Carlo study was conducted by generating 4,000 values of R/s at each sample size and counting how often the critical points were exceeded. The results (expressed as proportions of 4,000) are shown in Table 3. Sampling variation is inherent in Monte Carlo estimation and with this replication the standard errors of proportions are .008 for $\alpha = .5$, .007 for $\alpha = .3$, .005 for $\alpha = .1$, etc.

Table 3: *Relative Frequency of Exceeding Beta Points*

Nominal α	*n=20*	*n=40*	*n=60*	*n=80*	*n=100*
.5	.477	.472	.478	.464	.487
.4	.378	.373	.379	.372	.395
.3	.280	.285	.285	.275	.298
.2	.189	.194	.194	.188	.208
.1	.104	.106	.100	.100	.115
.05	.056	.059	.053	.055	.064
.01	.014	.013	.013	.018	.021

In interpreting the figures, it needs to be remembered that the simulations are independent between sample sizes, but all α points within a sample size are determined from the same 4,000 simulated R/s values. So all the α points in the n=100 column being a little higher than the corresponding figures in other columns is, at least partly, a sampling variation effect. Overall the "true" α is a little below the nominal for high α and a little above it for the 1 per cent tail point, but the Beta approximation is generally quite good enough to employ in obtaining "P values" and performing significance tests.[1] There is some indication that the overestimation at the 1 per cent point increases at high n. However, note that Table 2 showed the large deviation approximation to be good in this situation.

[1] Once again, upper tail rejection regions are assumed.

IV CONCLUDING REMARKS

The two approximations examined in this paper are simple to employ and, in reasonably sized samples, are vastly superior to the asymptotic approximation which, as Table 1 shows, is very poor. Neither of these new approximations is perfect, of course, although they complement each other. Other Beta type approximations are possible too. For example, instead of taking 2R/(ns) as the square root of a Beta, as in the previous section,

$$y = \frac{R/s-1}{n/2-1},$$

which lies between zero and one and becomes 2R/(ns) for large n, could be approximated by a Beta by equating first and second moments. This was examined by simulation, but only at high n and the 1 per cent level did it prove somewhat better than equating to the square root of a Beta. It was appreciably worse at low n. This suggests a more elaborate approximation of 2R/(ns) to Beta to the power m, with m to be determined. Then the expectation of 2R/(ns) and the approximate expectation of its square, which will be denoted c_m and c_{2m} respectively, would equate to the m th and 2m th moments of a Beta given by

$$\frac{\Gamma(p+q)\Gamma(p+m)}{\Gamma(p+q+m)\Gamma(p)}$$

and

$$\frac{\Gamma(p+q)\Gamma(p+2m)}{\Gamma(p+q+2m)\Gamma(p)},$$

respectively. The three parameters m, p and q cannot be estimated from the two known c_m and c_{2m}, but an approximation to the expectation of R/s cubed would lead to a c_{3m} and this could be equated to

$$\frac{\Gamma(p+q)\Gamma(p+3m)}{\Gamma(p+q+3m)\Gamma(p)},$$

giving solutions for all three parameters. Clearly an approximation to the expectation of R/s cubed can be obtained in the same way as (11) was by working from the asymptotic third central moment of R. A superficial examination suggested that m increases with n and that at n=20 the value should be less than .5, while at n=100 it should be closer to 1 than to .5, which ties in with the previous findings.

We have not taken this further, partly because the large deviation and two moment Beta approximations of Sections II and III seem jointly adequate for practical purposes, as Tables 2 and 3 show. The other reason is that, rather than seeking finer approximations in the simple normal sample context of this paper, future research would seem better directed to study of the robustness of the approximations to deviations from that context. If they are not reasonably robust, more computationally intensive approaches could perhaps be preferable. Izzeldin and Murphy (2000) have shown that the small sample distribution of the R/s statistic can be successfully bootstrapped using a variety of data generation processes.

REFERENCES

ANIS, A.A., and E.H. LLOYD, 1975. "Skew Inputs and the Hurst Effect", *Journal of Hydrology,* Vol. 26, pp. 39-53.

ANIS, A.A., and E.H. LLOYD, 1976. "The Expected Value of the Adjusted Rescaled Hurst Range of Independent Normal Summands", *Biometrika,* Vol. 63, pp. 111-116.

CONNIFFE, D., and J.E. SPENCER, 2000. "Approximating the Distribution of the Maximum Partial Sum of Normal Deviates", *Journal of Statistical Planning and Inference*, Vol. 88, Issue 1, pp. 19-27.

COX, D.R., and D.V. HINKLEY, 1974. *Theoretical Statistics,* London: Chapman and Hall.

FELLER, W., 1951. "The Asymptotic Distribution of the Range of Sums of Independent Random Variables", *Annals of Mathematical Statistics,* Vol. 22, pp. 427-432.

GEARY, R.C., 1933. "A General Expression for the Moments of Certain Symmetrical Functions of Normal Samples", *Biometrika,* Vol. 25, pp. 184-186.

HARRISON, M., and G. TREACY, 1997. "On the Small Sample Distribution of the R/S Statistic", *The Economic and Social Review*, Vol. 28, No. 4, pp.357-380.

HARRISON, M., and G. TREACY, 1998. *Testing for parameter instability using the R/S Statistic,* Technical Paper No. 21, Trinity Economic Papers, Dublin: Trinity College.

HURST, H.E., 1951. "Long Term Storage Capacity of Reservoirs", *Transactions of the American Society of Civil Engineers,* Vol. 116, pp. 770-799.

IZZELDIN, M., and A. MURPHY, 2000. "Bootstrapping the Small Sample Distribution of the R/s Statistic", Paper presented to the Annual Conference of the Irish Economic Association, Waterford.

JAMES, B., K.L. JAMES, and D. SIEGMUND, 1987. "Tests for a Change-point", *Biometrika*, Vol. 74, pp. 71-83.

KENDALL, M.G., and A. STUART, 1967. *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.

LLOYD E., 1981. "Stochastic Hydrology: An Introduction to Wet Statistics for Dry Statisticians", *Communications in Statistical Theory and Methods A,* Vol. 10, No. 15, pp. 1505-1522.

MANDELBROT, B., 1971. "When Can Price be Arbitraged? A Limit to the Validity of the Random Walk and Martingale Models", *Review of Economics and Statistics,* Vol. 53, pp. 225-236.

MANDELBROT, B., 1972. "Statistical Methodology for Non-periodic Cycles: from the Covariance to R/S Analysis", *Annals of Economic and Social Measurement,* Vol. 1, pp. 259-290.

PHIEN, H.N., A. ARBHABHIRAMA, and P. SUTABUTR, 1979. "The Water Storage Problem with Independent Normal Inflows", *Journal of the Science Society of Thailand*, Vol. 5, pp. 73-88.

PLOBERGER, W., and W. KRAMER, 1992. "The CUSUM Test with OLS Residuals", *Econometrica,* Vol. 60, pp. 271-285.

SHAZAM, 1997. *User's Reference Manual Version 8.0*, New York: McGraw-Hill.

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

SIEGMUND, D., 1986. Boundary Crossing Probabilities and Statistical Applications, *The Annals of Statistics*, Vol. 14, pp. 361-404.

SOLARI, M.E., and A.A. ANIS, 1957. "The Mean and Variance of the Maximum of the Adjusted Partial Sums of a Finite Number of Independent Normal Variates", *Annals of Mathematical Statistics,* Vol. 28, pp. 706-716.

SPITZER, F., 1956. "A Combinatorial Lemma and its Applications to Probability Theory", *Transactions of the American Mathematical Society,* Vol. 82, pp. 323-339.