THE ASYMPTOTIC DISTRIBUTION OF THE RANGE OF SUMS OF INDEPENDENT RANDOM VARIABLES

By William Feller

Princeton University

Summary. The asymptotic distribution of the range and normalized range of the sum of n independent variables is derived using the theory of Brownian motion.

1. Introduction. Let $[X_k]$ be a sequence of mutually independent random variables with a common distribution $V(x)$, and suppose that $E(X_k) = 0$, $\mathrm{Var}(X_k) = 1$. Put $S_n = X_1 + \cdots + X_n$ and let

$$M_n = \max [0, S_1, S_2, \cdots, S_n], \qquad m_n = \min [0, S_1, S_2, \cdots, S_n]. \tag{1.1}$$

The random variable

$$R_n = M_n - m_n \tag{1.2}$$

will be called *the range of the cumulative sums* S_n.

In applications[1] it is advantageous to modify this definition. One considers instead of the values of the sums S_k their deviations from the straight line joining the origin to the point (n, S_n). Thus we replace the random variables S_k by

$$S_k^* = S_k - kS_n/n \qquad (k = 1, \cdots, n) \tag{1.3}$$

and define the corresponding variables M_n^*, m_n^*, R_n^* in analogy with (1.1) and (1.2). The variable R_n^* will be called the *adjusted range of the cumulative sums* S_n.

The adjusted range has a greater sampling stability, but its main advantage is probably due to the fact that it eliminates the trend when $E(X_k) \neq 0$, so that it can be used even when the means do not vanish.

It is practically impossible to calculate the exact distribution of the ranges even for $n = 3$ and simple forms of the underlying distribution $V(x)$. Now the sums S_n are obviously asymptotically normally distributed, and therefore the asymptotic distribution of the ranges is independent of the form of $V(x)$. It suffices accordingly to consider the case where the variables X_k are normal. The sum S_n can then be considered as the value at time $t = n$ of a continuously changing normal variable $S(t)$ which is subject to a Bachelier-Wiener process (or ordinary diffusion). Since the sequence $[S_k]$ is a subsequence of the values assumed by $S(t)$, the range R_n is certainly not smaller than the range at time

[1] Cf. in particular Hurst [4]. A surprising statistical phenomenon discovered by Hurst is discussed at the end of Section 2. The author is indebted to Mr. G. W. Alexander of the State Rivers and Water Supply Commission, Melbourne, for drawing his attention to Hurst's paper and the interesting statistical problems connected with it.

$t = n$ of the variable $S(t)$, and it is clear that for large n the two ranges will be practically the same.

In Section 3 we shall find *the exact distribution of the range $R(t)$ of the continuous variable $S(t)$* [cf. (3.7)]. *One gets in particular*

$$\begin{aligned} E(R(n)) &= 2(2n/\pi)^{\frac{1}{2}} = 1.5958 \cdots n^{\frac{1}{2}}, \\ \mathrm{Var}(R(n)) &= 4n(\log 2 - 2/\pi) = 0.2181 \cdots n. \end{aligned} \tag{1.4}$$

These quantities are, asymptotically, the mean and varance of the range R_n.

For the *adjusted range* R_n^* we have to introduce the corresponding continuously changing variable

$$S^*(t) = S(t) - tS(T)/T \qquad (0 < t < T). \tag{1.5}$$

This variable appears more complicated than $S(t)$. Fortunately the stochastic process defined by (1.5) happens to be equivalent to a process which has been studied in an exceedingly elegant and simple manner by Doob in connection with his heuristic approach to the Kolmogorov-Smirnov theorems. Using Doob's results it is easy to obtain *the exact distribution of the adjusted range $R^*(T)$ for the continuously changing variable $S(t)$*. It is given in (4.3) and represents the desired asymptotic distribution of the adjusted range R_n^* for $n = T$. One gets in particular

$$\begin{aligned} E(R^*(T)) &= (T\pi/2)^{\frac{1}{2}} = 1.2533 \cdots T^{\frac{1}{2}}, \\ \mathrm{Var}(R^*(T)) &= \left(\frac{\pi^2}{6} - \frac{\pi}{2}\right) T = 0.07414 \cdots T. \end{aligned} \tag{1.6}$$

2. Discussion. A comparison of (1.4) and (1.6) shows that the adjusted range has the advantage of greater sampling stability.

In order to get an idea about the goodness of the approximations (1.4) and (1.6) we compare them with the exact values in the perhaps most unfavorable case, namely where each variable X_k assumes only the values ± 1, each with probability $\frac{1}{2}$. For $n = 6, 10, 12$ we get

(2.1)

	Exact value	*Approximation* (1.4)
$E(R_6)$	3.0625	3.909···
$\mathrm{Var}(R_6)$	1.18360	1.309···
$E(R_{10})$	4.1523···	5.046···
$\mathrm{Var}(R_{10})$	2.0872···	2.181···
$E(R_{12})$	4.6377···	5.528···
$\mathrm{Var}(R_{12})$	2.545 ···	2.617···

For the adjusted ranges (and the same Bernoulli variables) the corresponding figures are

(2.2)

	Exact value	*Approximation* (1.6)
$E(R_6^*)$	2	$3.070\cdots$
$\mathrm{Var}(R_6^*)$	$0.4396\cdots$	$0.445\cdots$
$E(R_{10}^*)$	$2.954\cdots$	$3.963\cdots$
$\mathrm{Var}(R_{10}^*)$	$0.5822\cdots$	$0.7414\cdots$

Considering the smallness of our n and the fact that the assumed distribution of the X_k is most unfavorable for our approximation, the above results appear surprisingly good. They also bear out the expectation that the ranges of the sums S_n should be smaller than those of the corresponding continuously varying variables $S(t)$.

If the model of cumulative sums of independent random variables applies to a particular type of empirical phenomena, then the observed ranges should, on the average, increase with the square root of the length T of the observational period. Now there is available a huge body of statistics concerning annual water levels of rivers and lakes all over the world. It has naturally been assumed that such levels could reasonably be treated as the cumulative effect of sums of random variables, but in an interesting paper [4] H. E. Hurst discovered puzzling systematic departures. In fact, Hurst has collected an impressively large statistical material relating to water levels and other phenomena which seems to bear out the contention that *the observed adjusted ranges do not increase, as expected, like the square root of the observational period T, but like a higher power T^c*. The most surprising feature is the stability of the observed values of the exponent c: it varies only from 0.69 to 0.80, with a mean of 0.729 and standard deviation 0.092. Within the several separate groups of phenonema the stability of c is even greater. Hurst himself has not attempted an explanation of his interesting discovery.

It is conceivable that the phenomenon can be explained probabilistically, starting from the assumption that the variables X_k are not independent, but that X_{n+1} depends only on the actual value of S_n. For example, a high lake level creates additional outlets for the outflow and this in practice means a restoring force towards the average size. Mathematically this would require treating the variables X_k as a Markov process. In theory the method presented in this paper applies to this more general case, but the simple ordinary diffusion equation would have to be replaced by a general Fokker-Planck equation, and the solution of the corresponding boundary value problem is not explicitly known. We are here confronted with a problem which is interesting from both a statistical and a mathematical point of view.

3. The range. We have to deal with the variable $S(t)$ of a Bachelier-Wiener process; this means that $S(t)$ is a normal variable with mean 0 and variance t

($t > 0$), and the increment $S(t + h) - S(t)$ is a normal variable with mean 0 and variance h which is independent of $S(t)$ (and the values $S(\tau)$ for $\tau < t$). For fixed $u > 0$ and $v > 0$ we require the probability $F(T; u, v)$ of the event

$$M(T) \leq v, m(T) \geq -u, \tag{3.1}$$

where $M(T) > 0$ and $m(T) < 0$ denote, respectively, the maximum and minimum of $S(t)$ for $0 \leq t \leq T$. The corresponding probability density is given by the mixed derivative

$$f(T; u, v) = F_{uv}(T; u, v), \tag{3.2}$$

and it is easily seen that the density function $\delta(T; r)$ of the range $R(T) = M(T) + |m(T)|$ is

$$\delta(T; r) = \int_0^r f(T; u, r - u)\, du. \tag{3.3}$$

To calculate $F(T; u, v)$ we start from the density function $w(t, x; u, v)$ of the event that simultaneously $S(t) = x$, $M(t) \leq v$, and $m(t) \geqq -u$. By the definition of these functions we have

$$F(T; u, v) = \int_{-u}^{v} w(T, x; u, v)\, dx, \tag{3.4}$$

so that the required density $\delta(T; r)$ follows from $w(T, x; u, v)$ by routine calculations. Now it is easily seen[2] that $w(t, x; u, v)$ is simply the fundamental solution of the ordinary diffusion equation $w_t = \frac{1}{2}w_{xx}$ for the interval $-u < x < v$ with the boundary conditions $w(t, x; u, v) = 0$ when $x = -u$ or $x = v$. One gets by the so-called method of images[3]

$$t^{\frac{1}{2}} w(t, x; u, v) = \sum_{k=-\infty}^{\infty} \phi\left(\frac{2ku + 2kv - x}{t^{\frac{1}{2}}}\right) - \sum_{k=-\infty}^{\infty} \phi\left(\frac{2ku + 2(k-1)v + x}{t^{\frac{1}{2}}}\right), \tag{3.5}$$

where $\phi(x)$ stands for the normal density function with zero mean and unit variance. Carrying out the indicated operations we find finally for *the density function of the range* $R(t)$

$$\delta(t; r) = 8 \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \phi\left(\frac{kr}{t^{\frac{1}{2}}}\right). \tag{3.6}$$

In this form it is not even obvious that the function is positive, and it is readily seen that the mean can not be obtained by termwise integration of

[2] The reasoning is substantially the same as in the case of discrete random walks (cf. [3], chap. 14).

[3] Cf. problem 5 on p. 304 of [3]. Formula (3.5) can be derived from the formula given there by the passage to the limit described in section 6 of chapter 14. Cf. also [5], p. 213.

(3.6). Fortunately this function is closely related to the distribution function $L(z)$ which occurs in the Kolmogorov-Smirnov theorem on empirical distribution functions. The distribution function $L(z)$ can be written in two equivalent forms[4]

$$\begin{aligned} L(z) &= 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 z^2) \\ &= (2\pi)^{\frac{1}{2}} z^{-1} \sum_{k=1}^{\infty} \exp(-(2k-1)^2 \pi^2/8z^2). \end{aligned} \tag{3.7}$$

Clearly

$$\delta(t; r) = (2/\pi)^{\frac{1}{2}} r^{-1} L'(r/(2t^{\frac{1}{2}})). \tag{3.8}$$

The second representation in (3.7) shows that $z^{-a}L(z) \to 0$ as $z \to 0$ for any a, and hence an integration by parts shows that

$$\int_0^{\infty} \delta(t; r)\, dr = 8\pi^{-2} \sum_{k=1}^{\infty} (2k-1)^{-2} = 1. \tag{3.9}$$

A similar procedure then leads to the formulas (1.4).

4. The adjusted range. We have now to find the range of the continuously changing variable defined by (1.5). It is clear that $S^*(t)$ is normally distributed with mean 0, variance $t(T-t)$, and $\mathrm{Cov}[S^*(s), S^*(t)] = s(T-t)/T$, for $0 < s < t < T$. Thus the stochastic process defined by (1.5) is, for the particular value $T = 1$, the process studied by Doob [1]. According to Doob a simple transformation permits one to reduce (1.5) to the ordinary Bachelier-Wiener process with the interval $0 < t < T$ going over into the entire interval $0 < t < \infty$. This actually simplifies matters inasmuch as the probabilities corresponding to (3.4) and (3.5) are no longer time dependent, so that the preceding boundary value problem for a partial differential equation is replaced by a simpler functional equation. At any rate, Doob's last equation furnishes us with the probability $F(T; u, v)$ that $S^*(t)$ is for $0 < t < T$ contained in the interval $(-u, v)$. We have[5]

$$\begin{aligned} F(T; u, v) &= 1 + e(u+v) \\ &\quad - \sum_{k=1}^{\infty} \{e(ku + (k-1)v) + e((k-1)u + kv) - e(ku + kv) \\ &\qquad - e((k-1)u + (k-1)v)\}, \end{aligned} \tag{4.1}$$

[4] Cf. formula (1.4) of [2], where however a factor 2 is missing in the exponent.

[5] Doob's formula looks simpler than (4.1), but the rearrangement (4.1) was necessary to make it possible to perform the required differentiations and integrations in the routine manner. (In the original form each term of the series contains a singular probability distribution along the axes, and formal manipulations lead into apparent contradictions. Also, Doob has $T = 1$.)

where we put for abbreviation

$$e(x) = \exp(-2x^2/T). \tag{4.2}$$

Formula (4.1) corresponds to (3.4), and it remains to perform the calculations indicated in (3.2) and (3.3). In this way we get for *the density function of the range* $R^*(T)$ *of the variable* $S^*(t)$

$$\delta(T; r) = re''(r) + \sum_{k=2}^{\infty} \{2k(k-1)[e'((k-1)r) - e'(kr)] + (k-1)^2 re''((k-1)r) + k^2 re''(kr)\}. \tag{4.3}$$

To see how the moments are calculated note, for example, that

$$\int_0^\infty r^2\delta(T; r)\,dr = \int_0^\infty r^3 e''(r)\,dr + \sum_{k=2}^{\infty}\left\{\left[\frac{2k}{(k-1)^2} - \frac{2k-2}{k^2}\right]\int_0^\infty r^2 e'(r)\,dr + \left[\frac{1}{(k-1)^2} + \frac{1}{k^2}\right]\int_0^\infty r^3 e'(r)\,dr\right\}. \tag{4.4}$$

But

$$\int_0^\infty r^3 e''(r)\,dr = -3\int_0^\infty r^2 e'(r)\,dr = 6\int_0^\infty re(r)\,dr = 3T/2,$$

and therefore

$$\int_0^\infty r^2\delta(T; r)\,dr = \frac{3}{2}T - \frac{1}{2}T\sum_{k=2}^{\infty}\left\{\frac{2k}{(k-1)^2} - \frac{(2k-2)}{k^2} - \frac{3}{(k-1)^2} - \frac{3}{k^2}\right\} = T\sum_{k=1}^{\infty} k^{-2} = \pi^2 T/6. \tag{4.5}$$

In this way formulas (1.6) are obtained.

REFERENCES

[1] J. L. Doob, "Heuristic approach to the Kolmogorov-Smirnov theorems," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 393–403.

[2] W. Feller, "On the Kolmogorov-Smirnov limit theorems for empirical distributions," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 177–189.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, John Wiley and Sons, 1950.

[4] H. E. Hurst, "Long-term storage capacity of reservoirs," *Proc. Amer. Soc. Civil Engrs.*, Vol. 76 (1950), separate no. 11, 30 pp.

[5] P. Lévy, *Processus stochastiques*, Gauthier-Villars, Paris, 1948.