The Annals of Mathematical Statistics
1972, Vol. 43, No. 1, 236-241

THE DIVISION OF A SEQUENCE OF RANDOM VARIABLES TO FORM TWO APPROXIMATELY EQUAL SUMS

By Aidan Sudbury and Peter Clifford

Bristol University

The finite sequence of n random variables $U_1 U_2 \cdots, U_n$ is divided into two complementary groups of random variables in one of 2^n ways. The random variables in each group are summed and the two sums are compared. Let $|S_n|$ be the minimum of the difference of the sums out of all the 2^n possible divisions. A lower bound to all sequences $\{\varepsilon_n\}$ such that $P\{|S_n| < \varepsilon_n\} \rightarrow 1$ as $n \rightarrow \infty$ is found in two cases:- $U_i = X_i$ $i = 1, 2 \cdots n$ and $U_i = X_i/\sum_{i=1}^n X_i$, $i = 1, 2 \cdots n$ where the X_i are independent and identically distributed random variables which have densities and satisfy certain regularity conditions.

The results lead to the solution of the particular problem of minimising the difference between two sums formed from segments of a fractured unit interval.

1. Introduction. The original version of this problem was proposed to one of the authors by Dr. M. Kanter at the University of Tel Aviv in 1969. To fix ideas we will first of all formulate the problem in its original form. The unit interval is broken into n segments by throwing down, randomly and independently, $n - 1$ uniformly distributed breaking points. The n segments are then reconstructed into two intervals of approximately equal length. We wish to investigate exactly how close an approximation can be made. In particular if $|S_n|$ is the minimum difference in length of the two reconstructed intervals we will find a lower bound to all sequences $\{\psi_n\}$ which have the property that $P\{|S_n| < \psi_n\} \rightarrow 1$ as $n \rightarrow \infty$. We write $G_n(\varepsilon) = P\{\min_\delta |\sum_1^n \delta_i X_i| < \varepsilon\}$, $\delta_i = \pm 1$, and prove the following:

Theorem. *If $\{X_i\}$ is a sequence of independent and identically distributed random variables such that each X_i has a finite third moment and the* νth *power of its characteristic function is integrable for some $\nu > 1$, then $\lim_{n\to\infty} G_n(\varepsilon_n) = 1$ for any sequence $\{\varepsilon_n\}$ such that $\varepsilon_n 2^n/(n)^{\frac{1}{2}} \rightarrow \infty$. Furthermore, if the sequence $\{\varepsilon_n\}$ is such that*

$$\lim_{n\to\infty} (2^n \varepsilon_n/\sigma(2n\pi)^{\frac{1}{2}}) = \omega\,, \quad \textit{then} \quad (\omega/(\omega + \tfrac{1}{2}))^2 < \lim_{n\to\infty} G_n(\varepsilon_n) < \omega\,.$$

Corollary. *If $\psi_n(n)^{\frac{1}{2}} 2^n \rightarrow \infty$ then*

$$\lim_{n\to\infty} P\{\min_\delta |\textstyle\sum_{i=1}^n \delta_i X_i/\sum_1^n X_i| < \psi_n\} = 1\,.$$

If $U_1, U_2, \cdots, U_n$ are the lengths of the broken segments we immediately have the well-known result $\mathscr{L}(U_1, U_2, \cdots, U_n) = \mathscr{L}(X_1/T, X_2/T, \cdots, X_n/T)$ where $T = \sum_{i=1}^n X_i$ and X_i, $i = 1, 2, \cdots, n$ are independent, identically and exponentially distributed random variables. The minimum difference $|S_n|$

Received December 22, 1970.

becomes $\min_\delta | \sum_{i=1}^n \delta_i X_i/T |$ where $\delta_1 = \pm 1, i = 1, 2, \cdots, n$, and the required result is a special case of the corollary.

Darling [1] has considered several other useful functionals on the lengths of the intervals formed on a line by random division, and Flatto and Konheim [3] have done much the same for the circle. The classical compendium of results in geometric probability is that of Kendall and Moran [4], and more recent results are presented in two review papers by Moran [5], [6].

2. The basic approach. Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed random variables. We wish to determine

$$G_n(\varepsilon) = P\{\min_\delta | \textstyle\sum_{i=1}^n \delta_i X_i | < \varepsilon\}$$

where $\delta_i = \pm 1, i = 1, \cdots n$.

Let $x_1, \cdots, x_n$ be a particular value of the sequence $X_1, \cdots, X_n$ and consider $T_n(x)$, the table of the 2^n possible values of $\sum_{i=1}^n \delta_i x_i$. Let $N_\varepsilon(x)$ be the number of entries in the table which lie in $(-\varepsilon, \varepsilon)$. Let

$$p_k = P\{N_\varepsilon(X) = 2k\}, \qquad k = 0, 1, \cdots 2^{n-1},$$

then

$$G_n(\varepsilon) = 1 - P\{N_\varepsilon(X) = 0\} = p_1 + p_2 + \cdots + p_{2^{n-1}} . \tag{1}$$

Instead of constructing the table we could toss a fair coin n times and obtain the n independent outcomes $I_1, \cdots, I_n$ where

$$P\{I_i = 1\} = P\{I_i = -1\} = \tfrac{1}{2} .$$

The sum $\sum_{i=1}^n I_i X_i$ will then have the same distribution as a single random sample from the 2^n entries in $T_n(X)$.

Let

$$F_1(\varepsilon) = P\{| \textstyle\sum_{i=1}^n I_i X_i | < \varepsilon\} ,$$

then, since it is the chance of selecting an element less than ε from $T_n(X)$,

$$\begin{aligned} F_1(\varepsilon) &= E_X\{N_\varepsilon(X)/2^n\} \\ &= \frac{1}{2^{(n-1)}}(p_1 + 2p + \cdots + 2^{n-1}p_{2^n-1}) . \end{aligned} \tag{2}$$

The first line of (2) tells us that the expected number of entries in $T_n(X)$ of size less than ε is $2^n F_1(\varepsilon)$.

We may also consider a series of trials in which we toss two fair coins n times each to give us the two sums $\sum I_i X_i$ and $\sum J_i X_i$ where

$$P\{J_i = 1\} = P\{J_i = -1\} = \tfrac{1}{2} .$$

The two sums will have the same joint distribution as a pair of random samples drawn with replacement from the table $T_n(X)$.

If follows that

$$(3)\qquad \begin{aligned} F_2(\varepsilon) &= P\{|\sum I_iX_i| < \varepsilon, |\sum J_iX_i| < \varepsilon\} \\ &= E\{(N_\varepsilon(X)/2^n)^2\} = \frac{1}{2^{2n-2}} \sum_{k=1}^{2^{n-1}} k^2 p_k \,. \end{aligned}$$

Using the functions defined above we may now determine an upper and lower bound for $G_n(\varepsilon)$. Except for small ε the upper bound is greater than 1 and the lower is negative, however they will be used to give the asymptotic behaviour as $\varepsilon \to 0$.

For any η

$$E\{(N_\varepsilon(X) - \eta)^2 | N_\varepsilon(X) > 0\} \geqq 0 \,,$$

and thus we have

$$-E\{(N_\varepsilon(X))^2\} + 2\eta E\{N_\varepsilon(X)\} \leqq \eta^2 P\{N_\varepsilon(X) > 0\} \,;$$

so using (1), (2) and (3) we may see that

$$(4)\qquad 2^{n-1}F_1(\varepsilon) \geqq G_n(\varepsilon) \geqq \eta^{-2}[2\eta 2^n F_1(\varepsilon) - 2^{2n}F_2(\varepsilon)] \,.$$

3. Approximations for large n. We shall approximate $P\{|\sum I_iX_i| < \varepsilon\}$ using the central limit theorem. If Y_i, $i = 1, \cdots n$ is a sequence of i. i. d. r. v.'s whose third moment μ_3 exists and such that the νth power of the characteristic function is integrable for some $\nu > 1$, then the density f_n of $Y_1 + \cdots + Y_n$ exists for $n > \nu$, and as $n \to \infty$ we have

$$f_n(x) - \frac{1}{\sigma(n)^{\frac{1}{2}}}\phi\left(\frac{x}{\sigma(n)^{\frac{1}{2}}}\right) = \frac{\mu_3}{6\sigma^5 n^{\frac{3}{2}}}\left(\frac{x^3}{\sigma^2 n} - 3x\right)\phi\left(\frac{x}{\sigma(n)^{\frac{1}{2}}}\right) + \frac{\lambda_{n(x)}}{n} \,,$$

where $\phi(x)$ is the standard normal density and $\lambda_n(x) \to 0$ as $n \to \infty$ uniformly in x. Feller [2] page 506.

If the X_i satisfy the above conditions, then $\mu_3 = 0$ for the I_iX_i, and letting f_n now stand for the density of $\sum I_iX_i$ we have that

$$(5)\qquad f_n(x) = \frac{1}{\sigma(n)^{\frac{1}{2}}}\phi\left(\frac{x}{\sigma(n)^{\frac{1}{2}}}\right) + o\left(\frac{1}{n}\right) .$$

By the mean value theorem there exists an x in $(-\varepsilon, \varepsilon)$ such that

$$(6)\qquad \begin{aligned} F_1(\varepsilon) &= P\{|\sum I_iX_i| < \varepsilon\} = 2 \in f_n(x) \\ &= \frac{2\varepsilon}{\sigma(2n\pi)^{\frac{1}{2}}} + \varepsilon o\left(\frac{1}{n}\right) , \end{aligned}$$

using (5), where we are only considering ε's less than some fixed positive number ε_0.

We now wish to find an approximation for $F_2(\varepsilon)$. We note that $\sum (I_i + J_i)X_i/2$ and $\sum (I_i - J_i)X_i/2$ never involve the same variables, and are of the form $\sum_{k=1}^m \delta_{i_k} X_{i_k}$. Thus $P\{|\sum I_iX_i| < \varepsilon, |\sum J_iX_i| < \varepsilon\}$ is the sum over all possible subsets of $X_1, \cdots, X_n$, which we shall designate by $X_{i_1}, \cdots, X_{i_m}$, of the probabilities that $\sum (I_i + J_i)X_i/2$ involve exactly $X_{i_1}, \cdots, X_{i_m}$ and that

$$|I_{i_1}X_{i_1} + \cdots + I_{i_m}X_{i_m} + \cdots + I_{i_n}X_{i_n}| < \varepsilon ,$$
$$|I_{i_1}X_{i_1} + \cdots + I_{i_m}X_{i_m} - I_{i_{m+1}}X_{i_{m+1}} - \cdots - I_{i_n}X_{i_n}| < \varepsilon .$$

Since the X_i's are identically distributed

$$F_2(\varepsilon) = \frac{1}{2^n} \textstyle\sum_{m=0}^{n} \binom{n}{m} P\{|\sum_1^n I_iX_i| < \varepsilon, |\sum_1^m I_iX_i - \sum_{m+1}^n I_iX_i| < \varepsilon\} ,$$
$$= \frac{2F_1(\varepsilon)}{2^n} + \frac{1}{2^n} \textstyle\sum_{m=1}^{n-1} \binom{n}{m} P\{|\sum_1^n I_iX_i| < \varepsilon, |\sum_1^m I_iX_i - \sum_{m+1}^n I_iX_i| < \varepsilon\} .$$

We now find an approximation for the second term. Now $\sum_1^m I_iX_i$ and $\sum_{m+1}^n I_iX_i$ are independent random variables with densities f_m and f_{n-m}; thus by the mean value theorem there are values x_m and y_m such that $|x_m + y_m| < \varepsilon$ and $|x_m - y_m| < \varepsilon$, and such that the second term, in the expression above equals

$$2\varepsilon^2 \textstyle\sum_{m=1}^{n-1} \binom{n}{m} \frac{1}{2^n} f_m(x_m) f_{n-m}(y_m) .$$

Using (5) this becomes, for $\varepsilon < \varepsilon_0$,

$$(7) \qquad \frac{2\varepsilon^2}{2\pi\sigma^2} \textstyle\sum_{m=1}^{n-1} \binom{n}{m} \frac{1}{2^n} [m^{-\frac{1}{2}} + om^{-1}]\left[(n-m)^{-\frac{1}{2}} + o\left(\frac{1}{n-m}\right)\right] .$$

We will next show that

$$(8) \qquad \lim_{n\to\infty} \textstyle\sum_{m=1}^{n-1} \binom{n}{m} \frac{n}{2^n}\left(m^{-\frac{1}{2}} + \frac{\lambda_m}{m}\right)\left((n-m)^{-\frac{1}{2}} + \frac{\lambda_{n-m}}{n-m}\right) = 2 ,$$

where $\lambda_n \to 0$ as $n \to \infty$. Let $\lambda_n{}^* = \max_{1\leqq m\leqq n-1} |\lambda_m|$, then

$$(9) \qquad \left|(m(n-m))^{-\frac{1}{2}} + \frac{\lambda_m}{m}(n-m)^{-\frac{1}{2}} + \frac{\lambda_{n-m}}{n-m} m^{-\frac{1}{2}} + \frac{\lambda_{n-m}\lambda_m}{m(n-m)}\right|$$
$$< (n-1)^{-\frac{1}{2}} + \frac{2\lambda_n{}^*}{(n-1)^{\frac{1}{2}}} + \frac{\lambda_n{}^{*2}}{n-1}$$

for $1 \leqq m \leqq n-1$.

Let $\delta > 0$ be arbitrarily small and $N(\delta)$ be such that, for all $n \geqq N(\delta)$,

$$(10) \qquad \left(1 + 2\lambda_n{}^* + \frac{\lambda_n{}^{*2}}{(n-1)^{\frac{1}{2}}}\right)\frac{n}{(n-1)^{\frac{1}{2}}}\frac{1}{4n^{\frac{3}{4}}} < \frac{\delta}{2} , \qquad \text{and furthermore}$$

if $|m/n - \frac{1}{2}| < n^{-1/8}$, then

$$(11) \quad |(m(1-m/n)/n)^{-\frac{1}{2}} - 2| < \frac{\delta}{8} , \quad \left|\frac{\lambda_m}{m}\frac{n}{(n-m)^{\frac{1}{2}}}\right| < \frac{\delta}{8} , \text{ and } \left|\frac{n\lambda_{n-m}\lambda_m}{m(n-m)}\right| < \frac{\delta}{8} .$$

Let $B(m) = \sum_{k=0}^{m} \binom{n}{k} \frac{1}{2} n$, so that $B(m)$ is the distribution function of a binomial distribution with variance $n/4$. We consider the integral

$$\int_{1 \leqq m \leqq n-1} \left| n \left(m^{-\frac{1}{2}} + \frac{\lambda_m}{m} \right) \left((n-m)^{-\frac{1}{2}} + \frac{\lambda_{n-m}}{n-m} \right) - 2 \right| dB(m) , \tag{12}$$

and divide it into two parts, the first of which is over the region $[1 \leqq m \leqq n-1] \cap \{|m/n - \frac{1}{2}| < n^{-1/8}\}$. From (11) we may see that this is $< \delta/2$. The second is over the rest of the region $1 \leqq m \leqq n-1$. By the Chebychev inequality the probability of lying in this region is $\leqq 4n^{-\frac{3}{4}}$, and thus, using (9) and (10), this part of the integral is also $< \delta/2$.

So given any $\delta > 0$, the expression on the left of (8) differs from

$$\int_{1 \leqq m \leqq n-1} 2 dB(m) = 2 - \frac{1}{2^{n-2}} ,$$

by less than δ for all $n > N(\delta)$. Thus

$$F_2(\varepsilon) = \frac{2}{2^n} F_1(\varepsilon) + \frac{\varepsilon^2}{n\pi\sigma^2} [2 + \delta(n)] , \tag{13}$$

where $\delta(n) \to 0$ as $n \to \infty$ provided ε remains bounded.

4. The limiting behavior of $G_n(\varepsilon)$. Using equations (4), (6) and (13), we have that for all η

$$G_n(\varepsilon) \geqq \frac{1}{\eta^2} \left[\frac{2^{n+2}\eta\varepsilon}{\sigma(2n\pi)^{\frac{1}{2}}} - \frac{2^{n+2}\varepsilon}{\sigma(2n\pi)^{\frac{1}{2}}} - \frac{2^{2n+1}\varepsilon^2}{n\pi\sigma^2} \right]$$
$$\left[\varepsilon 2^{n+1} \left(\frac{1}{\eta} - \frac{1}{\eta^2} \right) - 2^{2n} \frac{\varepsilon^2}{\eta^2} \right] o \left(\frac{1}{n} \right) .$$

For a given η we maximise the first of this expression by putting

$$\frac{\varepsilon . 2^{n+1}}{\sigma(2n\pi)^{\frac{1}{2}}} = \eta - 1 . \tag{14}$$

We then have

$$G_n \left[\frac{(n-1)\sigma(2n\pi)^{\frac{1}{2}}}{2^{n+1}} \right] \geqq \left(1 - \frac{2}{\eta} + \frac{1}{\eta^2} \right) + \left(\frac{n-1}{n} \right)^2 (2n)^{\frac{1}{2}} - (\pi)^{\frac{1}{2}} \sigma n/2) o \left(\frac{1}{n} \right) .$$

So for any sequence $\{\varepsilon_n\}$ converging to zero such that $\varepsilon_n 2^{n+1}(\sigma^2 2n\pi)^{-\frac{1}{2}} \to \infty$ or equivalently $\varepsilon_n 2^n (n)^{-\frac{1}{2}} \to \infty$ we have

$$\lim_{n \to \infty} G_n(\varepsilon_n) = 1 . \tag{15}$$

Also, since $G_n(\varepsilon) < 2^{n-1} F_1(\varepsilon)$, we may say that for any set of values γ_n such that

$$\lim_{n \to \infty} \frac{2^n \gamma_n}{\sigma(2n\pi)^{\frac{1}{2}}} = \omega ,$$

$$(\omega/(\omega + \tfrac{1}{2}))^2 < \lim_{n\to\infty} G_n(\gamma_n) < \omega \,. \tag{16}$$

Let us now consider random variables of the form

$$Y_{i,n} = X_i/\textstyle\sum_{i=1}^n X_i\,, \quad i = 1, \cdots, n; n = 1, 2, \cdots .$$

We will assume that the X_i's satisfy the conditions leading to (15).

Consider

$$\begin{aligned} P\{\min_{\delta_i} |\textstyle\sum_{i=1}^n \delta_i Y_{i,n}| < \psi_n\} &= P\{\min_{\delta_i} |\textstyle\sum_{i=1}^n \delta_i X_i| < n\psi_n\,|\bar{X}_n|\} \\ &\geqq P\{\min_{\delta_i} |\textstyle\sum_{i=1}^n \delta_i X_i| < (|\mu| - \delta)n\psi_n\} - P\{|\bar{X}_n| < |\mu| - \delta\} \end{aligned}$$

where $\mu = E(X_i)$ and $\bar{X}_n = \sum_{i=1}^n X_i/n$.

Let $\delta = 1/n^{\frac{1}{4}}$ then using (15) and the law of large numbers we see that

$$P\{\min_\delta |\textstyle\sum_{i=1}^n \delta_i Y_{i,n}| < \psi_n\} \to 1$$

as $n \to \infty$ and $n\psi_n \to 0$, provided $\psi_n(n)^{\frac{1}{4}}2^n \to \infty$, but it is clear that only the last condition is necessary.

5. Conclusion. Equations (15) and (16) for the asymptotic behavior can hardly be improved on. (16) puts a lower limit on the ε_n's such that $\lim_{n\to\infty} G_n(\varepsilon_n) = 1$, and (15) states that any sequence tending to zero 'infinitely more slowly' than this value will have that property. However, in any particular case where the distribution of the random variables is known, the asymptotic behavior gives us no help in estimating the probability for a given ε and n. This estimate can only be derived by calculating the various terms that have been shown to tend to zero as n tends to infinity. In most cases the error in (8) predominates, and for any numerical estimates it would clearly be best to calculate the LHS directly. When the distribution function is known, the error due to approximating by the central limit theorem may also be calculated.

For values of ε not very small it appears that $G_n(\varepsilon)$ will be close to 1, and the most interesting thing to calculate is the expected number of elements of the form $\sum \delta_i X_i$ which are of size less than ε. From (2) we may see that this is $2^n F_1(\varepsilon)$, which in many cases is not difficult to work out.

REFERENCES

[1] Darling, D.A. (1953). On a class of problems related to the random division of an interval. *Ann. Math. Statist.* **24** 239–273.

[2] Feller, W. (1966). *An Introduction to Probability Theory and its Applications* **2**. Wiley, New York.

[3] Flatto, L. and Kinheim, G. (1962). The random division of an interval and the random covering of a circle. *SIAM Review* **4** No. 3.

[4] Kendall, M.G. and Moran, P.A.P. (1963). Geometrical Probability. Griffin, London.

[5] Moran, P.A.P. (1966). A note on recent research in Geometrical Probability. *J. Appl. Probability* **3** 453–563.

[6] Moran, P.A.P. (1969). A second note on recent research in Geometrical Probability. *Advances Appl. Probability* **1** 73–90.