SIMPLE RANDOM WALK AND RANK ORDER STATISTICS[1]

BY MEYER DWASS

Northwestern University

1. Introduction. Suppose $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ are $2n$ independent random variables, each having the same *continuous* cdf. Let the $2n$ random variables be arranged in increasing order, and the values of the X's replaced by (-1)'s and the values of Y's replaced by 1's. This sequence of (-1)'s and 1's is called the sequence of *rank order indicators*. There are $\binom{2n}{n}$ equally likely sequences of rank order indicators. A random variable which is a function of the X's and Y's only through these rank order indicators is called a *rank order statistic*. Often rank order statistics are defined in terms of the empirical cdf's F_n and G_n:

$$F_n(t) = (\text{number of } X\text{'s} \leqq t)/n,$$

$$G_n(t) = (\text{number of } Y\text{'s} \leqq t)/n.$$

Some examples of rank order statistics are

$$D_n^+ = \sup_{-\infty<t<\infty} (F_n(t) - G_n(t))$$
(one-sided Kolmogorov-Smirnov statistic),

$$D_n = \sup_{-\infty<t<\infty} |F_n(t) - G_n(t)|$$
(two-sided Kolmogorov-Smirnov statistic)

With every sequence of rank order indicators one can associate a "random·walk" graph in a familiar way. (See Figure 1.) As indicated in Figure 1, this random walk moves from (0, 0) to $(2n, 0)$ performing n upward steps and n downward steps, with the $\binom{2n}{n}$ possible paths being equally likely. Gnedenko and his coworkers succeeded in determining the distributions of various rank order statistics including D_n^+ and D_n by combinatoric analyses of the above described random walk. This is now a much used technique and accounts for the methodology in the papers [2], [4], [5], [6], [7], [8]. The above random walk consists of *nonindependent* steps. The purpose of this paper is to provide an alternate method based on ordinary *simple random walk* which has *independent* steps. We believe this new method gives simple and unified proofs for the distributions previously derived and for new distributions as well.

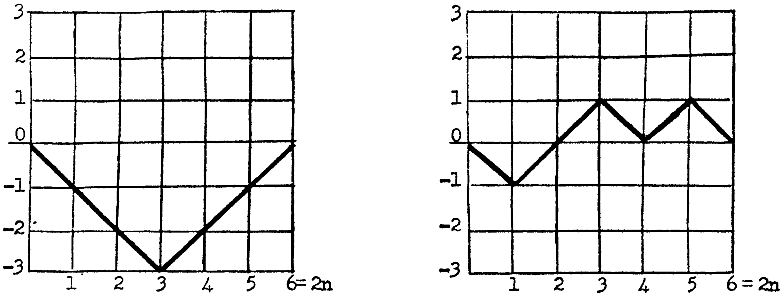
2. The method. Suppose that $W_1, W_2, \cdots$ are independent and identically distributed random variables with distribution given by

$$W_i = \begin{cases} 1, & \text{with probability } p, \\ -1, & \text{with probability } 1 - p = q. \end{cases}$$

The following lemma is easy to verify.

Received 6 December 1967.

[1] Research sponsored by USAF Contract AF 49(638)-1480.



This graph represents the sequence $-1, -1, -1, 1, 1, 1$.

This graph represents the sequence $-1, 1, 1, -1, 1, -1$.

FIG. 1. A downward step denotes -1 and an upward step denotes 1. The graph starts at $(0, 0)$ and ends at $(2n, 0)$. There are $\binom{2n}{n}$ possible paths.

LEMMA 1. *For any p in $(0, 1)$ the conditional distribution of $W_1, \cdots, W_{2n}$ given that $W_1 + \cdots + W_{2n} = 0$ assigns equal probabilities to each of the $\binom{2n}{n}$ possible sequences of n 1's and n (-1)'s. In other words, the distribution is exactly that of the rank order indicators described in Section* 1.

If $p < \frac{1}{2}$ then the simple random walk, $W_1, W_1 + W_2, W_1 + W_2 + W_3, \cdots$ is *transient*. With probability 1 there are only finitely many returns to the origin. (In other words, $W_1 + \cdots + W_{2n} = 0$ for only finitely many values of n. Of course, a return to the origin cannot take place for an odd index.) Let T be the time of last return of the random walk to the origin. That is, T is the largest value of $2n$ for which $W_1 + \cdots + W_{2n} = 0$. The following assumption is important in what follows.

DEFINITION. Let U be a function defined on the random walk. We say that U satisfies *Assumption* A if the value of U is completely determined by $W_1, \cdots, W_T$, whenever $T > 0$.

Closely related to Lemma 1 is the following:

LEMMA 2 (a). *The conditional distribution of $W_1, \cdots, W_T$, given that $T = 2n$, assigns equal probabilities to each of the $\binom{2n}{n}$ possible sequences of n (-1)'s and n 1's. If U satisfies Assumption A, the conditional distribution of U given that $T = 2n$ is exactly that of a rank order statistic.*

(b). *Conversely, suppose U_n is a rank order statistic defined for every $n = 1, 2, \cdots$. Then there is a function U satisfying Assumption A, such that the conditional distribution of U given that $T = 2n$ is exactly the distribution of U_n.*

PROOF. Part (a) follows immediately from Lemma 1. To prove part (b) simply define

$$U(W_1, \cdots, W_T) = U_n(W_1, \cdots, W_{2n})$$

when $T = 2n > 0$. When $T = 0$, U can be defined arbitrarily. This definition makes sense since U_n, being a rank order statistic, is defined for every sequence of n (-1)'s and n 1's.

NOTATION. In what follows, we shall use the dual notation U, U_n as suggested in Lemma 2. If U satisfies Assumption A, U_n is a rank order statistic, U will denote the function satisfying Assumption A defined in the proof of Lemma 2.

The main tool which we use to determine distributions of rank order statistics is the following:

THEOREM. *Suppose U_n is a rank order statistic for every n and U is the related function satisfying Assumption A. Define*

$$E(U) = h(p), \qquad 0 \leqq p < \tfrac{1}{2}.$$

Then the following power series (in powers of $p(1 - p) = pq$) is valid for $0 \leqq p < \frac{1}{2}$.

$$h(p)/(1 - 2p) = \textstyle\sum_{n=0}^{\infty} E(U_n)\binom{2n}{n}(pq)^n. \tag{2.1}$$

PROOF. For $p < \frac{1}{2}$,

$$h(p) = E(U) = \textstyle\sum_{n=0}^{\infty} E(U \mid T = 2n)P(T = 2n) = \sum_{n=0}^{\infty} E(U_n)P(T = 2n),$$

by Lemma 2. Since the probability that the simple random walk never returns to the origin is $1 - 2p$ (see Appendix (2)), it follows that $P(T = 2n) = \binom{2n}{n}\cdot(pq)^n(1 - 2p)$. Hence $h(p)/(1 - 2p) = \sum_{n=0}^{\infty} E(U_n)\binom{2n}{n}(pq)^n$.

REMARKS. (a) If φ is a function defined over the possible values of U then $\varphi(U_n)$ is also a rank order statistic. In particular if φ is the set indicator function of A then $E(\varphi(U_n)) = P(U_n \text{ in } A)$. In the applications of the theorem we shall let the symbols U, U_n represent $\varphi(U)$, $\varphi(U_n)$ for the various versions of φ that may be convenient to the problem at hand, without further comment.

(b) The usefulness of the theorem in determining distributions of rank order statistics depends on the ease with which one can explicitly evaluate h and then determine the power series expansion $h(p)/(1 - 2p) = \sum_{n=0}^{\infty} a_n(pq)^n$. Once such a power series expansion is available, then since the a_n's are uniquely determined, we immediately read off the relationship

$$E(U_n) = a_n/\binom{2n}{n}. \tag{2.2}$$

EXAMPLE. Let F_n, G_n be as defined in Section 1. Suppose that $Z_1 < Z_2 < \cdots < Z_{2n}$ are the ordered values of the combined set $X_1, \cdots, X_n, Y_1, \cdots, Y_n$. (Define $Z_0 = -\infty$.) Now define the rank order statistic U_n as follows:

$$U_n = \text{number of indices } 0 \leqq i \leqq 2n \text{ for which } F_n(Z_i) = G_n(Z_i).$$

(This can be roughly described as "the number of times the two empirical cdf's are equal to each other.") The related function U is then the number of times that the random walk visits the origin. Explicitly (with $W_0 \equiv 0$), U = number of indices i for which $W_0 + \cdots + W_i = 0, i = 0, 1, 2, \cdots, P(U > k) = (2p)^k$. (See Appendix (3).) Hence, identifying $h(p)$ as $P(U > k)$, we have

$$\begin{aligned} h(p)/(1 - 2p) &= (2p)^k/(1 - 2p) \\ &= 2^k \textstyle\sum_{n=k}^{\infty} \binom{2n-k}{n-k}(pq)^n. \quad \text{(See Appendix (14).)} \end{aligned}$$

Hence, by (2.2) $P(U_n > k) = 2^k\binom{2n-k}{n-k}/\binom{2n}{n}, n = k, k + 1, \cdots$.

In the remainder of this paper we have made a selection of derivations of distributions which appear in the literature. We provide alternate proofs based on (2.1). This list is not meant to be encyclopedic. We also present some distributions which appear to us to be new. Sometimes in the literature after a distribution has been determined by complicated combinatoric means, a "generating function" is determined by even more complicated means. This generating function turns out to be essentially (2.1). It does not seem to have been noticed however that (2.1) can be directly and easily evaluated using simple random walk and then the distribution determined from the generating function. See [7], [8] and the proofs of VIIIa in this paper for an illustration of this point.

3. List of rank order statistics. Recall that $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ are $2n$ independent and identically distributed random variables, with F_n the empirical cdf of the X's and G_n the empirical cdf of the Y's. Let $Z_1 < Z_2 < \cdots < Z_{2n}$ be the ordered values of the combined set of X's and Y's and let $Z_0 = -\infty$. Define

$$n[F_n(u) - G_n(u)] = H_n(u), \qquad -\infty < u < \infty.$$

(Notice that by our definition of cdf, $H_n(u) = H_n(u+)$.)

The following is the list of rank order statistics whose distributions will be derived. (Figure 2 illustrates these definitions in terms of the random walk diagram.)

I. N_n, *the number of times* F_n *equals* G_n.

N_n = number of indices i for which $H_n(Z_i) = 0, i = 0, 1, 2, \cdots, 2n$.

II. N_n^+, N_n^-, *the positive and negative sojourns.*

Let $0 < i_1 < i_2 < \cdots < i_{N_n} = 2n$ be the indices for which $H_n(Z_i) = 0$. If $H_n(Z_i) > 0$ for $i_{j-1} < i < i_j$, we say the jth sojourn is positive, and if $H_n(Z_i) < 0$ for $i_{j-1} < i < i_j$, we say the jth sojourn is negative. (Use the convention

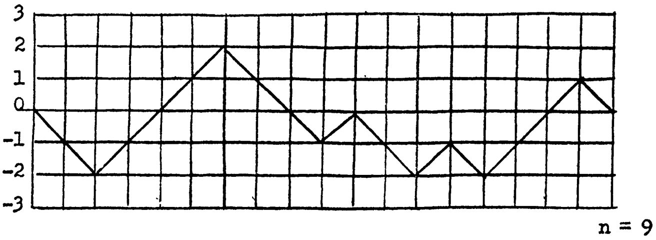


FIG. 2. The above is the graph of the rank order indicators −1, −1, 1, 1, 1, 1, −1, −1 −1, 1, −1, −1, 1, −1, 1, 1, 1, −1,

$$N_n = 6$$
$$N_n^+ = 2, \quad N_n^- = 3$$
$$N_n(1) = 3, \; N_n(2) = 1, \; N_n(3) = 0$$
$$N_n^+(0) = 2, \; N_n^+(1) = 1$$
$$N_n^*(0) = 3, \; N_n^*(1) = 2$$
$$D_n^+ = 2, \quad D_n^- = 2, \quad D_n = 2$$
$$Q_n = 1$$
$$Q_{n,1} = 4, \quad Q_{n,2} = 8, \quad Q_{n,3} = 10, \; Q_{n,4} = 16$$
$$R_n^+ = 6$$
$$L_n = 6.$$

that $i_0 = 0$. Notice that since $H(Z_i) = 0$ only for even indices i then the sets $i_{j-1} < i < i_j$ are not empty.) Let

$$N_n^+ = \text{number of positive sojourns and}$$
$$N_n^- = \text{number of negative sojourns.}$$

III. $N_n(r)$, *number of visits to height* r.

$$N_n(r) = \text{number of indices } i \text{ for which } H(Z_i) = r, r = 0, 1, \cdots .$$

IV. $N_n^+(r)$, *upcrossings of* r.

$$N_n^+(r) = \text{number of indices } i \text{ for which } H(Z_i) = r + 1 \quad \text{and } H(Z_{i-1}) = r, r = 0, 1, \cdots .$$

V. $N_n^*(r)$, *number of crossings of* r.

$N_n^*(r) =$ number of indices i for which

$$H_n(Z_i) = r \text{ and } [r - H_n(Z_{i-1})]\,[r - H_n(Z_{i+1})] < 0, \quad r = 0, 1, \cdots .$$

(For $r = 0$, adopt the convention that a crossing automatically is counted for $i = 0$ and $i = 2n$.)

VI. D_n^+, D_n^-, *the one-sided maximum deviations.*

$$D_n^+ = \max_{-\infty<u<\infty} H_n(u) = \max_{0\leqq i\leqq 2n} H_n(Z_i),$$
$$D_n^- = \max_{-\infty<u<\infty} -H_n(u) = \max_{0\leqq i\leqq 2n} -H_n(Z_i).$$

VII. D_n, *the two sided maximum deviation.*

$$D_n = \max_{-\infty<u<\infty} |H_n(u)| = \max_{0\leqq i\leqq 2n} |H_n(Z_i)| = \max D_n^+, D_n^-.$$

VIII. Q_n, *the number of times* D_n^+ *is achieved.*

$$Q_n = \text{number of indices } i \text{ for which } H_n(Z_i) = D_n^+.$$

IX. $Q_{n,k}$, *position of* kth *zero.*

$Q_{n,k}$ equals the index i for which $H_n(Z_i) = 0, 0 < i$, for the kth time.

(This is defined only when $N_n \geqq k$.)

X. R_n^+, *index where* D_n^+ *is first achieved.*

$$R_n^+ = \text{smallest } i \text{ such that } H_n(Z_i) = D_n^+, \text{ if } D_n^+ > 0,$$
$$= 0, \text{ if } D_n^+ = 0.$$

XI. L_n, *length of positive sojourns.*

Recall the notation in II above. If the jth sojourn between i_{j-1} and i_j is positive the contribution to L_n is $i_j - i_{j-1}$. L_n is the sum of all such contributions over $[0, 2n]$.

4. Distribution results.

I. $P(N_n > k) = 2^k\binom{2n-k}{n-k}/\binom{2n}{n}$, $k = 1, 2, \cdots, n$.

II. $P(N_n^+ \geqq k) = P(N_n^- \geqq k) = \binom{2n}{n-k}/\binom{2n}{n}, k = 0, 1, \cdots, n.$

III. $P(N_n(r) > k) = 2^k\binom{2n-k}{n-k-r}/\binom{2n}{n}, k = 0, 1, \cdots, n.$

IV. $P(N_n^+(r) \geqq k) = \binom{2n}{n-k-r}/\binom{2n}{n}, k = 0, 1, \cdots, n)$. (Theorem 1 of [6])

V. If $r > 0, k = 1, 2, \cdots, P(N_n(r) \geqq 2k) = \binom{2n}{n-r-2k+1}/\binom{2n}{n}$. (Equivalent to Theorem 2.1, [2].)

If $r = 0, k = 1, 2, \cdots, P(N_n(0) \geqq k) = 2\binom{2n-1}{n-k-1}/\binom{2n}{n}$. (Equivalent to Theorem 1.1′, [2])

VI(a). $P(D_n^+ \geqq k) = P(D_n^- \geqq k) = \binom{2n}{n-k}/\binom{2n}{n}, k = 0, 1, \cdots, n$. [4]

VI(b). $P(D_n^+ < k, D_n^- < r) = 1 - \sum_{i=1}^{\infty} [\binom{2n}{n-ik-(i-1)r} + \binom{2n}{n-(i-1)k-ir} - 2\binom{2n}{n-ik-ir}]/\binom{2n}{n}, k = 1, \cdots, n; r = 1, \cdots, n$. (Theorem 1 of [5].) (There are only a finite number of non-zero terms in this summation.)

VII. $P(D_n < k) = 1 - 2\sum_{i=1}^{\infty} [\binom{2n}{n-(2i-1)k} - \binom{2n}{n-2ik}]/\binom{2n}{n}, k = 1, \cdots, n$. [4]

VIII(a). $P(D_n^+ \geqq k, Q_n \geqq r) = \binom{2n-r+1}{n-k-r+1}/\binom{2n}{n}, k = 0, 1, \cdots; r = 1, 2, \cdots.$

VIII(b).$P(Q_n \geqq \ ) = \binom{2n-r+1}{n-r+1}/\binom{2n}{n}, r = 1, \cdots, n.$

IX(a). If $r > 0$,

$$\begin{aligned} P(Q_{n,k} &= 2i, N_n = k + r) \\ &= 2^{k+r}[k/(2i - k)]\binom{2i-k}{i-k}[r/(2n - 2i - r)]\binom{2n-2i-r}{n-i-r}/\binom{2n}{n}, \\ &\quad i = k, \cdots, n - r. \end{aligned}$$

(Obviously, if $N_n = k$ then $Q_{n,k} = 2n$.)

IX(b). $P(Q_{n,k} = 2i, N_n \geqq k) = 2^k[k/(2i - k)]\binom{2i-k}{i-k}\binom{2n-2i}{n-i}/\binom{2n}{n}, i = k, \cdots, n.$

X(a).
$$\begin{aligned} P(D_n^+ &= k, R_n^+ = r) \\ &= [k(k + 1)/r(2n - r + 1)]\tfrac{1}{2}\binom{r}{r+k}\binom{2n-r+1}{n-\frac{1}{2}(r+k)}/\binom{2n}{n}, \text{ if } r + k \\ &\quad \text{is even, } k > 0. \end{aligned}$$

(If $k = 0$, then $R_n^+ = 0$.) (This is Theorem 1 of [7].)

X(b). $\frac{1}{2}(R_n^+ + D_n^+)$ is uniformly distributed over $0, 1, \cdots, n$.

$$P(\tfrac{1}{2}(R_n + D_n^+) = k) = 1/(n + 1), \qquad k = 0, 1, \cdots, n.$$

XI. $L_n/2$ is uniformly distributed over $0, 1, \cdots, n$.

$$P(L_n/2 = k) = 1/(n + 1), k = 0, 1, \cdots, n. \text{ [1]}$$

Proofs of Distribution Results. Recall that we are using the dual notation U, U_n. If U_n is a rank order statistic, then U is the function defined on the simple random walk (srw) as defined in Lemma 2. Let $W_0 = 0$ and $W_1, W_2, \cdots$ be independent random variables with

$$\begin{aligned} W_i &= 1 \text{ with probability } p < \tfrac{1}{2}, \\ &= -1 \text{ with probability } 1 - p = q. \end{aligned}$$

Define $S_n = W_0 + \cdots + W_n$, $n = 0, 1, \cdots$.

I. This was already proved in the example of Section 2.

II. Let N^+ denote the number of positive sojourns of the srw and N the number of returns to the origin. It is easy to verify that the conditional distribution of N^+ given that $N^+ = r$ is binomial with parameters $r, \frac{1}{2} = p/2p$. ($2p$ is the probability of ever returning to the origin. See Appendix (2).) Hence

$$P(N^+ = k) = \sum_0^\infty \binom{r}{k}(\tfrac{1}{2})^r(2p)^r(1 - 2p)$$

($(2p)^r(1 - 2p)$ is $P(N = r + 1)$. See Appendix (3).)

Hence $P(N^+ = k) = (1 - 2p)p^k/q^{k+1}$ and $P(N^+ \geqq k) = (p/q)^k$. Hence $h(p)/(1 - 2p) = (p/q)^k/(1 - 2p) = \sum_n \binom{2n}{n-k}(pq)^n$. (See Appendix (15).) Hence $P(N_n^+ \geqq k) = \binom{2n}{n-k}/\binom{2n}{n}$, by (2.2).

III. Let $N(r)$ be the number of indices i for which $S_i = r$. Then $P(N(r) > k) = (p/q)^r(2p)^k$ by Appendix (6) and (3).

$$\begin{aligned} h(p)/(1 - 2p) &= (p/q)^r(2p)^k/(1 - 2p) \\ &= 2^k p^{k+2r}/(pq)^r(1 - 2p) = 2^k \sum_{n=k+2r}^\infty \binom{2n-k-2r}{n-k-2r}(pq)^{n-r} \\ &= 2^k \sum_{n=k}^\infty \binom{2n-k}{n-k-r}(pq)^n. \end{aligned}$$

Hence $P(N_n(r) > k) = 2^k\binom{2n-k}{n-k-r}/\binom{2n}{n}$.

IV. Let $N^+(r)$ be the number of indices i for which $S_{i-1} = r$, $S_i = r + 1$, $i = 1, 2, \cdots$. By Appendix (12), $P(N^+(r) \geqq k) = (p/q)^{k+r}$. Hence

$$\begin{aligned} h(p)/(1 - 2p) &= (p/q)^{k+r}/(1 - 2p) \\ &= \sum_n \binom{2n}{n-k-r}(pq)^n, \qquad \text{by Appendix (15).} \end{aligned}$$

Therefore $P(N_n^+(r) \geqq k) = \binom{2n}{n-k-r}/\binom{2n}{n}$.

V. Let $N^*(r)$ be the number of indices for which srw crosses height r, that is, $N^*(r)$ equals the number of i for which

$$S_{i-1} < r = S_i < S_{i+1} \quad \text{or} \quad S_{i-1} > r = S_i > S_{i+1}, \quad 0 \leqq i \leqq T^{\cdot}$$

Then $P(N^*(r) \geqq 2k) = (p/q)^{r+2k-1}$, if $r > 0$. ($(p/q)^{r+1}\cdot 1$ is the probability of reaching $r + 1$ and then going down to $r - 1$ (first 2 crossings); $(p/q)^2\cdot 1$ is the probability of going from $r - 1$ to $r + 1$ and back to $r - 1$ again (next 2 crossings), etc.) Also

$$P(N^*(0) \geqq k) = 2p^{k+1}/q^k.$$

($P(N^*(0) \geqq 1) = p\cdot 1\cdot(p/q) + q(p/q)^2\cdot 1 = 2p^2/q$, etc.) Hence for $r > 0$,

$$h(p)(1 - 2p)^{-1} = (p/q)^{r+2k-1}(1 - 2p)^{-1} = \sum_n \binom{2n}{n-r-2k+1}(pq)^n$$

(Appendix (15))

and $$P(N_n^*(r) \geqq k) = \binom{2n}{n-r-2k+1}/\binom{2n}{n}.$$

For $r = 0$,

$$h(p)(1 - 2p)^{-1} = (2p^{k+1}/q^k)(1 - 2p)^{-1} = 2(p^{2k+1}/(pq)^k)(1 - 2p)^{-1}$$

$$= 2\sum_n \binom{2n-2k-1}{n-2k-1}(pq)^{n-k}$$
$$= 2\sum_n \binom{2n-1}{n-k-1}(pq)^n.$$

Hence $P(N^*(0) \geqq k) = 2\binom{2n-1}{n-k-1}/\binom{2n}{n}$.

VI(a). Let $D^+ = \max(0, S_1, S_2, \cdots)$. By Appendix (6), $P(D^+ \geqq k) = (p/q)^k$. Hence $h(p)/(1 - 2p) = (p/q)^k/(1 - 2p)$ which equals $\sum_{n=k}^{\infty} \binom{2n}{n-k} \cdot (pq)^n$ by Appendix (15). Hence $P(D_n{}^+ \geqq k) = \binom{2n}{n-k}/\binom{2n}{n}$.

VI(b). By Appendix (10)

$$P(-r < S_n < k, n = 1, \cdots, T; \text{u}\, T = 0)$$
$$= [1 - (p/q)^k][1 - (p/q)^r][1 - (p/q)^{k+r}]^{-1}.$$

Hence

$$h(p)(1 - 2p)^{-1} = [1 - (p/q)^k - (p/q)^r + (p/q)^{k+r}]$$
$$\cdot \sum_{1=0}^{\infty} (p/q)^{(k+r)i}(1 - 2p)^{-1}$$
$$= [1 - \sum_{i=1}^{\infty} (p/q)^{ik+(i-1)r} - \sum (p/q)^{(i-1)k+ir}$$
$$+ 2\sum_{i=1}^{\infty} (p/q)^{i(k+r)i}](1 - 2p)^{-1}$$
$$= \sum_n a_n(pq)^n$$

where $a_n = 1 - \sum_i \binom{2n}{n-ik-(i-1)r} + \binom{2n}{n-(i-1)k-ir} - 2\binom{2n}{n-ik-ii}$ by Appendix (15).

VII. Simply set $r = k$ in VI(b).

VIII. Let $D^+ = \max(0, S_1, S_2, \cdots)$, Q = number of indices i for which $S_i = D^+$. By Appendix (11), $P(Q \geqq r, D^+ \geqq k) = (p/q)^k p^{r-1}$.

$$h(p)/(1 - 2p) = (p/q)^k p^{r-1}/(1 - 2p) = p^{2k+r-1}/(pq)^k(1 - 2p)$$
$$= \sum_{n=2k}^{\infty} \binom{2n-2k-r+1}{n-2k-r+1}(pq)^{n-k} \qquad \text{(by Appendix (14))}$$
$$= \sum_{n=k}^{\infty} \binom{2n-r+1}{n-k-r+1}(pq)^n.$$

Hence $P(Q_n \geqq r, D_n{}^+ \geqq k) = \binom{2n-r+1}{n-k-r+1}/\binom{2n}{n}$. For $P(Q_n \geqq r)$ set $k = 0$.

IX(a). As in the proof of II, let N be the total number of returns of srw to 0 and let Q_k be the position of the kth 0. Then

$$\sum_i P(Q_k = i, N = k + r)t^i = [1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k(2p)^r(1 - 2p)$$

by Appendix (1) and (2). Hence

$$h(p)/(1 - 2p) = [1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k(2p)^r$$
$$= 2^{-k}\cdot[1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k \cdot 2^{-r}\cdot[1 - (1 - 4pq)^{\frac{1}{2}}]^r 2^{k+r}$$
$$= 2^{k+r}[k\sum_m m^{-1}\binom{2m-k-1}{m-k}(pq)^m t^{2m}]\,[r\sum_j \binom{2j-r-1}{j-1}(pq)^j],$$

by Appendix (16).

$\binom{2n}{n}P(Q_{n,k} = 2i, N_n = k + r)$ is the coefficient of $t^{2i}(pq)^n$ which is easily checked to be

$$2^{k+r}[k/(2i - k)]\binom{2i-k}{i-k}[r/(2n - 2i - r)]\binom{2n-2i-r}{n-i-r}.$$

IX(b). $\sum P(Q_k = i, N \geqq k)t^i = [1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k$.

$$h(p)/(1 - 2p) = [1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k/(1 - 4pq)^{\frac{1}{2}}$$
$$= [2^k k \sum_m m^{-1}\binom{2m-k-1}{m-k}(pq)^m t^{2m}]\,[\sum_j \binom{2j}{j}(pq)^j].$$

$\binom{2n}{n}P(Q_{n,k} = 2i, N_n \geqq k)$ is the coefficient of $t^i(pq)^n$ which is easily checked to be $[2^k k/(2i - k)]\binom{2i-k}{i-k}\binom{2n-2i}{n-i}$.

X(a). Let R^+ be the index at which the maximum of srw is first achieved. $R^+ =$ smallest i for which $S_i = D^+$, $i = 0, 1, 2, \cdots$.

$$E(t^{R+}; D^+ = k) = (2qt)^{-k}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k(1 - p/q),$$

by Appendix (7) and (5).

$$E(t^{R+}; D^+ = k)(1 - 2p)^{-1} = 2^{-k}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k\, p^{k+1}/(pq)^{k+1}t^k$$
$$= 2^{-k}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k$$
$$\cdot 2^{-(k+1)}[1 - (1 - 4pq)^{\frac{1}{2}}]^{k+1}/(pq)^{k+1}t^k$$
$$= [\sum_i (2i - k)^{-1}\binom{2i-k}{i-k}(pq)^i]$$
$$\cdot[\sum_j (2j - k - 1)^{-1}\binom{2j-k-1}{j-k-1}(pq)^j]\,[(pq)^{k+1}t^k]^{-1}.$$

$\binom{2n}{n}P(R_n{}^+ = r, D_n{}^+ = k)$ is the coefficient of $t^n(pq)^n$ in the above. This coefficient is easily checked to be

$$k(k + 1)[r(2n - r + 1)]^{-1}\binom{r}{(r+k)/2}\binom{2n-r+1}{n-((r+k)/2)} \text{ where } r + k \text{ is even.}$$

X(b). Let $V_1, V_2, \cdots, V_{D^+}$ be the times it takes srw to go successive steps upward, (0 if $D^+ = 0$). Then

$$(V_1 + 1) + (V_2 + 1) + \cdots + (V_{D^+} + 1) = R^+ + D^+.$$

The generating function of $V_i + 1$ is $t[1 - (1 - 4pqt^2)^{\frac{1}{2}}](2qt)^{-1}$ by Appendix (4). Hence

$$E(t^{R^++D^+}) = \sum_{k=0}^{\infty} [1 - (1 - 4pqt^2)^{\frac{1}{2}}]^k[(2q)^{-1}]^k(1 - (p/q))$$
$$= (1 - 2p)\{q - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]\}^{-1}.$$

Referring to the proof of XI below, $E(t^{R^++D^+}) = E(t^L)$. Hence $R_n{}^+ + D_n{}^+$ and L_n have the same distribution.

XI. Let L be the length of all positive sojourns in srw. By Appendix (13)

$$E(t^L)/(1 - 2p) = \{q - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]\}^{-1}$$
$$= \tfrac{1}{2}[1 - (1 - 4pq)^{\frac{1}{2}}] - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}][pq(1 - t^2)]^{-1}$$

(by some elementary algebra)

$$= \sum_m \sum_n (2n + 1)^{-1}\binom{2n+1}{n}(pq)^n t^{2m}$$

$$- \sum_m \sum_n (2n + 1)^{-1}\binom{2n+1}{n}(pq)^n t^{2(n+m+1)}$$

(by Appendix (16)).

$\binom{2n}{n}P(L_n = k)$ equals the coefficient of $(pq)^n t^{2k}$ in $E(t^L)/(1 - 2p)$. Hence

$$P(L_n = k) = (n + 1)^{-1} \text{ if } k = 0, 2, \cdots, 2n$$
$$= 0 \text{ if } k > 2n.$$

APPENDICES

Appendix on Simple Random Walk (srw)

All the facts that we need either appear in Feller [3], or are easily derived from elementary considerations. The following list covers what is needed in this paper. We suppose as before that $p < \frac{1}{2}$. As above, $S_n = W_0 + \cdots + W_n$, $n = 0, 1, \cdots$.

(1) The generating function for return time to the origin is

$$1 - (1 - 4pqt^2)^{\frac{1}{2}} = \psi(t). \qquad \text{[3], p. 257}$$

(2) The probability of ever returning to the origin is $\psi(1) = 2p$.

(3) The probability of being at the origin more than k times is $(2p)^k$, $k = 0, 1, 2, \cdots$. (This counts an initial visit at time 0.)

(4) The generating function for the time to reach 1 is $\psi(t)/2qt$. [3], p. 255

(5) The probability of ever reaching 1 is $\psi(1)/2q = p/q$.

(6) The probability of ever reaching k is $(p/q)^k$, $k = 1, 2, \cdots$.

(7) The generating function for the time to reach k is $[\psi(t)/2qt]^k$, $k = 1, 2, \cdots$. (This follows from (4).)

(8) The generating function for the time to reach k, with $-r$ not being reached by that time is

$$(\psi(t)/2qt)^k[1 - (\psi(t)/2t)^{2r}(pq)^{-r}]/[1 - (\psi(t)/2t)^{2(k+r)}(pq)^{-(k+r)}],$$

$k = 1, 2, \cdots; r = 1, 2, \cdots$. (This is (4.11) p. 320 Feller [3].)

(9) The probability of hitting k before hitting $-r$ is

$$(\psi(1)/2q)^k[1 - (\psi(1)/2)^{2r}(pq)^{-r}]/[1 - (\psi(1)/2)^{2(k+r)}(pq)^{-(k+r)}]$$
$$= (p/q)^k[1 - (p/q)^r]/[1 - (p/q)^{k+r}].$$

(This follows from (8).)

(10) $P(-r < S_n < k, n = 1, \cdots, T; \text{u}\, T = 0) = [1 - (p/q)^k]\,[1 - (p/q)^r]\cdot[1 - (p/q)^{k+r}]^{-1}$.

Proof. The required probability is 1 minus the probability of hitting k before hitting $-r$ or hitting $-r$ before hitting k and then returning to the origin. The first probability is $(p/q)^k[1 - (p/q)^r]\,[1 - (p/q)^{k+r}]^{-1}$ by (9). The second probability is

$$\{1 - (p/q)^k[1 - (p/q)^r]\,[1 - (p/q)^{k+r}]^{-1}\} \times (p/q)^r.$$

The sum of these two probabilities is

$$1 - [1 - (p/q)^k][1 - (p/q)^r][1 - (p/q)^{k+r}]^{-1}.$$

(11) Let $D^+ = \max(0, S_1, S_2, \cdots)$. Let $Q =$ number of indices i for which $S_i = D^+$.

$$P(Q = r, D^+ = k) = (p/q)^k p^{r-1}(1 - 2p);$$

$$P(Q \geqq r) = p^{r-1};$$

$$P(Q \geqq r, D^+ \geqq k) = (p/q)^k p^{r-1}.$$

PROOF. $P(Q = r, D^+ = k) = (p/q)^k(q \cdot p/q)^{r-1}(1 - 2p)$. $(p/q)^k$ is the probability of reaching height k. Never to exceed k but to reach it again $r - 1$ times has probability of $(qp/q)^{r-1}(1 - 2p)$. Q and D^+ are independent. The second and third assertions follow from summation of the first.

(12) Let $N^+(r)$ equal the number of times srw upcrosses height r, $r = 0, 1, \cdots$. Specifically, $N^+(r) =$ number of indices i for which $S_{i-1} = r$, $S_i = r + 1$.

$$P(N^+(r) \geqq k) = (p/q)^{k+r}.$$

PROOF. First observe that $P(N^+(0) \geqq k) = (p/q)^k$. $N(0) \geqq k$ means srw reaches 1 and returns to 0 successively at least k times. This event has probability $(p/q \cdot 1)^k$. $(p/q)^r$ is the probability of reaching r, and $(p/q)^k$ is the probability of then upcrossing r k times. Hence $P(N^+(r) \geqq k) = (p/q)^{k+r}$.

(13) Suppose i_1 and i_2 are two successive return times to the origin of srw. If $0 < S_i$, $i_1 < i < i_2$, the interval $[i_1, i_2]$ is called a *positive* sojourn and if $S_i < 0$, $i_1 < i < i_2$ it is called a negative sojourn. The length of the sojourn is $i_2 - i_1$. Let $L =$ the sum of lengths of all positive sojourns. Then

$$E(t^L) = (1 - 2p)\{q - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]\}^{-1}.$$

PROOF. The generating function for the positive contribution to any one sojourn is

$$pt[1 - (1 - 4pqt^2)^{\frac{1}{2}}](2pt)^{-1} + q(p/q) = \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}] + p.$$

(If the first step is positive then the contribution to the generating function is the first term by (4); if the first term is negative the contribution is $q(p/q)$ the probability of starting down and then returning.)

$$E(t^L) = \textstyle\sum_{k=0}^{\infty} \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}] + p)^k(1 - 2p)$$

$$= (1 - 2p)\{q - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{\frac{1}{2}}]\}^{-1}.$$

Appendix on power series

The following power series expansions in powers of pq are all valid for $0 \leqq p < \frac{1}{2}$.

(14) $p^k/(1 - 2p) = \sum_{n=k}^{\infty} \binom{2n-k}{n-k}(pq)^n$.

PROOF. Let $W_1, W_2, \cdots$ be variables in terms of which the srw is defined.

$$P(W_1 = W_2 = \cdots = W_k = 1) = p^k$$
$$= \textstyle\sum_{n=k}^{\infty} P(W_1 = W_2 = \cdots = W_k = 1 \mid T = 2n)\binom{2n}{n}(pq)^n(1 - 2p),$$

as in the proof of the Theorem in Section 2. But

$$P(W_1 = W_2 = \cdots = W_k = 1 \mid T = 2n) = \binom{2n-k}{n-k}/\binom{2n}{n}$$

since this is simply the probability that a sequence of n 1's and $n(-1)$'s starts off with k 1's, when all $\binom{2n}{n}$ sequences are equally likely.

(15) $(p/q)^k/(1 - 2p) = \sum_{n=k}^{\infty} \binom{2n}{n-k}(pq)^n$.

PROOF.

$$(p/q)^k = p^{2k}/(pq)^k = \textstyle\sum_{n=2k}^{\infty} \binom{2n-2k}{n-2k}(pq)^{n-k} (\text{by } (14))$$
$$= \textstyle\sum_{n=k}^{\infty} \binom{2n}{n-k}(pq)^n.$$

(16) $p^k = \frac{1}{2}[1 - (1 - 4pq)^{\frac{1}{2}}]^k = k\sum_{n=k}^{\infty} (2n - k)^{-1}\binom{2n-k}{n-k}(pq)^n$.

PROOF. That $p = \frac{1}{2}[1 - (1 - 4pq)^{\frac{1}{2}}]$ is a matter of elementary algebra. The mapping $pq = t$ is $1 - 1$ of $[0, \frac{1}{2})$ onto $[0, \frac{1}{4})$ with inverse $p = \frac{1}{2}[1 - (1 - 4t)^{\frac{1}{2}}]$. The transformed version of the expansion (14) with $k - 1$ instead of k is $\frac{1}{2}[1 - (1 - 4t)^{\frac{1}{2}}]^{k-1}/(1 - 4t)^{\frac{1}{2}} = \sum_{n=k-1}^{\infty} \binom{2n-k+1}{n-k+1}t^n$. Integrating both sides with respect to t and determining the constant of integration from $t = 0$

$$\tfrac{1}{2}[1 - (1 - 4t)^{\frac{1}{2}}]^k = k\textstyle\sum_{n=k-1}^{\infty} (n + 1)^{-1}\binom{2n-k+1}{n-k+1}t^{n+1}$$
$$= k\textstyle\sum_{n=k}^{\infty} (2n - k)^{-1}\binom{2n-k}{n-k}t^n.$$

Expressing this in terms of p gives the required result.

REFERENCES

[1] CHUNG, K. L. and FELLER, W. (1949). Fluctuations in coin tossing. *Proc. Nat. Acad. Sci. USA* **35** 605–608.

[2] CSÁKI, E. and VINCZE, I. (1961). On some problems connected with the Galton test. *Publ. Math. Inst. Hungarian Acad. Sci. Ser. A* **6** 97–109.

[3] FELLER, WILLIAM (1957). *An Introduction to Probability Theory and Its Applications* (2nd Ed.) Wiley, New York.

[4] GNEDENKO, B. V. and KOROLYUK, V. S. (1951). On the maximum discrepancy between two empirical distributions. (Russian) *Dokl. Akad. Nauk SSSR* **80** 525–528. (*Sel. Transl. Math. Statist. Prob.* 1961). 13–22.

[5] GNEDENKO, B. V. and RVACEVA, E. L. (1952). On a problem of the comparison of two empirical distributions. (Russian) *Dokl. Akad. Nauk SSSR.* **82** 513–516. (*Sel. Transl. Math. Statist. Prob.* (1961) 69–72.)

[6] MIHALEVIC, V. S. (1952). On the mutual disposition of two empirical distribution functions. (Russian) *Dokl. Akad. Nauk SSSR.* **85** 485–488. (*Sel. Transl. Math. Statist. Prob.* 1961, pp. 63–68)

[7] VINCZE, I. (1957). Einige zweidimensionale Verteilungs- und Grenzverteilungssätze in der Theorie der geordneten Stechproben. *Publ. Math. Inst. Hungarian Acad. Sci.* **2** 183–203.

[8] VINCZE, I. (1959). On some joint distributions and some joint limiting distributions in the theory of order statistics, II. *Publ. Math. Inst. Hungarian Acad. Sci.* **4** 29–41.