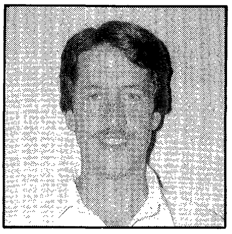
The Longest Run of Heads

Mark F. Schilling



Mark F. Schilling is Associate Professor at California State University, Northridge. He received his B.A. and M.A. in mathematics at the University of California at San Diego and his doctorate was earned in statistics at the University of California at Berkeley in 1979 under the supervision of Peter J. Bickel. Schilling was employed at the University of Southern California prior to his appointment at C.S.U. Northridge.

Dr. Schilling's research interests include statistical methods for multidimensional data and the probabilistic behavior of repetitive sequences. His hobbies include sports (and statistics), boomerang flying, music, and hiking.

The two sequences shown below each purportedly represent the results of 200 tosses of a fair coin. One of these is an actual sequence obtained from coin tossing, while the other sequence is artificial. Can you decide, in sixty seconds or less, which of the sequences is more likely to have arisen from actual coin tossing and which one is the imposter?

Sequence #1

T HHHHT T T T HHHHT HHHHHHHHT T THHT T HHHHHT T T T T HHT HHT HHH T
T T HT T HHHHT HT T T HT T T HHT T T T HHHHHHT T T HHT T HHHT HHHHHT T T T
T HT T T HHT T HT T HHT T T HHT T T HH THHT HHT T T T T HHT HHHHHHT HT HT T
HT HT T HHHT T HHT HT HHHHHHHHT THT T HHHT HHT T HT T T T T T HHHT HHH

Sequence #2

T HT HT T T HT T T T T HT HT T T HT T HHHT HHT HT HT HT T T T HHT T HHT T HHH T
HHHT T HHHT T T HHHT HHHHT T T HT HT HHHHT HT T T HHHT HHT HT T T HHT H
HHT HHHHT T HT HHT HHHT T T HT HHHT HHT T T HHHT T T T HHHT HT HHHHT H
T T HHT T T T HT HT HT T HT HHT T HT T THT T T T HHHHT HT HHHT T HHHHHT HH

The above challenge is based on a classroom experiment originally performed by Révész [**14**]. The class is divided into two groups. In the first group, each student is instructed to toss a coin 200 times and record the resulting sequence of heads and tails. Each student in the second group is merely to write down a sequence of heads and tails that the student believes is a reasonable *simulation* of 200 tosses of a fair coin. Given the combined results of the two groups, Révész claims that the students can be classified back into their original groups with a surprising degree of accuracy by means of a very simple criterion: In students' simulated patterns, the longest run of consecutive heads or consecutive tails is almost invariably *too short* relative to that which tends to arise from actual coin tossing.

The real coin tossing sequence above is #1, which has a longest run of eight heads (twice), while the longest run found in Sequence #2 is only five heads long. Before reading on, you may wish to conjecture answers to the following questions: What *is* a reasonable value for the length of the longest run of heads in n tosses of a fair coin? What about the length of the longest run of either heads *or* tails?

Curiosity about the above phenomenon has led me to conduct essentially the same experiment in courses in introductory probability theory; the resulting percentage of correct classifications has averaged around 85%. (Interestingly, those persons who have managed a successful deception by submitting a simulated sequence containing a long run have generally turned out later to be among my best students.) The fact that one can easily and in a matter of minutes separate the two groups quite well stimulates considerable student interest and provides a splendid topic for illustrating some important facets of probability theory, including recursion arguments, asymptotic analysis and the concept of limiting distributions, while at the same time strikingly driving home the message that human beings make rather poor randomization devices.

We begin by developing simple recursion formulas that generate the exact distribution of the longest run of heads, both for a fair coin and for a coin with probability of heads $p \in (0,1)$. Several curious features of head run distributions are then explored.

The Exact Distribution of the Longest Run

If a fair coin is flipped, say, three times, we can easily list all possible sequences:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT

and accordingly derive the exact distribution of the longest head run:

longest head run	*probability*
0	1/8
1	4/8
2	2/8
3	1/8

The expected length of the longest head run is 11/8. The probability histogram for the above distribution is shown in Figure 1, which contains for each value of x a rectangle centered at x whose height is the corresponding probability $P(x)$ of a run of length x.

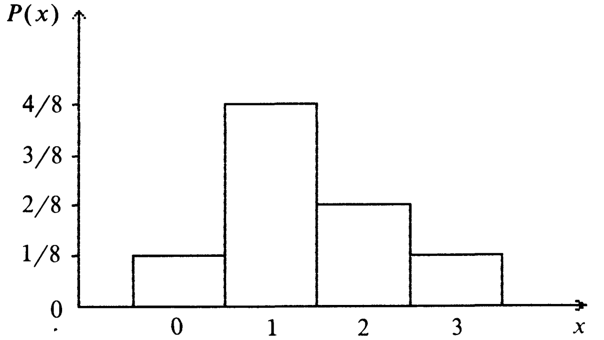


Figure 1
The distribution of the longest head run in three tosses of a fair coin

For n as small as six, however, it is quite laborious to compute the exact distribution of the longest head run by enumerating all cases; when $n = 200$, the staggering number 2^{200} of points in the sample space makes the 'sledgehammer' approach inaccessible even to large computers. The situation is complicated even further when the coin is not a fair coin, for it is no longer the case that each possible sequence has the same probability. Some finesse is clearly called for.

The case for a fair coin. Consider n independent tosses of a fair coin, and let R_n represent the length of the longest run of heads. The stochastic behavior of the longest head run can be described in terms of its probability distribution function, but it turns out to be much easier to deal instead with its *cumulative* distribution function

$$F_n(x) = P(R_n \le x).$$

Let $A_n(x)$ be the number of sequences of length n in which the longest run of heads does not exceed x. Clearly, $F_n(x) = 2^{-n}A_n(x)$, but how can we compute $A_n(x)$? The key is to partition the set of favorable outcomes (sequences) according to the number of heads, if any, that occur before the first tail. This leads to a simple recursive formula for $A_n(x)$.

To see how this works, consider the case in which the longest head run consists of three heads or fewer. If $n \le 3$ then clearly $A_n(3) = 2^n$ since any outcome is a favorable one. For $n > 3$, each favorable sequence begins with either T, HT, HHT, or HHHT and is followed by a string having no more than three consecutive heads. Thus

$$A_n(3) = A_{n-1}(3) + A_{n-2}(3) + A_{n-3}(3) + A_{n-4}(3) \quad \text{for } n > 3.$$

Using the recursion, the values of $A_n(3)$ can easily be computed:

n	0	1	2	3	4	5	6	7	8	$\cdots$
$A_n(3)$	1	2	4	8	15	29	56	108	208	$\cdots$

Thus for, say, $n = 8$ tosses of a fair coin, the probability is $208/2^8 = 0.8125$ that the longest head run has length no greater than 3. In the general case we obtain

$$A_n(x) = \begin{cases} \sum_{j=0}^{x} A_{n-1-j}(x) & \text{for } n > x; \\ 2^n & \text{for } n \le x. \end{cases} \tag{1}$$

Note that for $n = 1, 2, 3, \ldots$, the number $A_n(1)$ of sequences of length n that contain no two consecutive heads is the $(n + 2)$nd Fibonacci number.

The longest run of heads or tails. It is a simple matter to apply the partitioning argument above to obtain the distribution of the longest run of pure heads *or* pure tails for a fair coin. Let R'_n be the length of the longest such run and let $B_n(x)$ be the number of strings of length n for which R'_n is less than or equal to x. We then have $B_n(0) = 0$ and

$$B_n(x) = 2A_{n-1}(x - 1) \quad \text{for } x \ge 1. \tag{2}$$

To justify (2), observe that when proceeding through a sequence of coin tosses, a new run begins precisely when the outcome of the latest toss is different from that of the preceding one. The roles of H and T in (1) are now played by S and D, respectively, where S represents the event that an adjacent pair of coin tosses have the same outcome and D represents the event that they are different. The result then follows from these considerations: (i) n tosses result in $n-1$ adjacent pairs, (ii) the outcome of the first toss is irrelevant, and (iii) a string of $x-1$ consecutive S's is necessary and sufficient for a run of length x.. The example below shows one of the strings that contributes to $B_{10}(3)$:

T H T T T H T T H H

D D S S D D S D S

Letting $F_n'(x) = P(R_n' \leq x)$, we obtain easily from (2) that $F_n'(x) = F_{n-1}(x-1)$. This says that the distribution of the longest run of heads *or* tails for a fair coin is simply the distribution of the longest run of heads alone for a sequence containing one fewer coin toss, shifted to the right by one. For example, the chance that the longest head or tail run in 1000 tosses is, say, of length twelve is exactly the same as the chance that the longest head run in 999 tosses is eleven long. Since the distribution of longest runs is not greatly affected by one coin toss unless n is very small, the implication of (2) is that for n tosses of a fair coin the longest run of heads *or* tails, statistically speaking, tends to be about one longer than the longest run of heads alone.

Biased coins. Now consider the situation in which the probability of heads p can take any value in $(0,1)$. How does this affect the length of the longest head run and the longest run of heads or tails?

It is again possible to obtain a recursive result, but now it is necessary to refine the combinatorial analysis which was used for a fair coin by considering the *total* number of heads, k, in the sequence in addition to the length of the longest head run, since strings with different numbers of heads will have different probabilities of occurrence when $p \neq 1/2$.

Let $C_n^{(k)}(x)$ be the number of strings of length n in which exactly k heads occur, but no more than x of these occur consecutively. The cumulative distribution of the longest run then can be expressed as

$$F_n(x) = \sum_{k=0}^{n} C_n^{(k)}(x) p^k q^{n-k}, \tag{3}$$

where $q = 1 - p$ is the probability of tails. Observe that

$$A_n(x) = C_n^{(0)}(x) + C_n^{(1)}(x) + \cdots + C_n^{(n)}(x).$$

Let us again consider first the special case $x = 3$. Note that for $k \leq 3$, $C_n^{(k)}(3)$ is the binomial coefficient $\binom{n}{k}$, and that for $3 < k = n$, $C_n^{(k)}(3) = 0$. In the interesting case where $3 < k < n$, each of the $C_n^{(k)}(3)$ sequences begins with either T, HT, HHT, or HHHT and ends with a string having no more than 3 consecutive heads and a total of either k, $k-1$, $k-2$ or $k-3$ heads, respectively. Thus we get the recursion

$$C_n^{(k)}(3) = C_{n-1}^{(k)}(3) + C_{n-2}^{(k-1)}(3) + C_{n-3}^{(k-2)}(3) + C_{n-4}^{(k-3)}(3).$$

k \ n	0	1	2	3	4	5	6	7	8
8									0
7								0	0
6							0	1	10
5						0	2	12	40
4					0	3	12	31	65
3				1	4	10	20	35	56
2			1	3	6	10	15	21	28
1		1	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1	1	1

Figure 2
Values of $C_n^{(k)}(3)$ for $n \le 8$

Figure 2 displays the values of $C_n^{(k)}(3)$ for $n \le 8$. The first four rows of the figure ($k = 0, 1, 2, 3$) are part of Pascal's triangle. Entries above that are computed by taking diagonal sums of four entries from the rows and columns below and to the left. The 'hockey stick' illustrates the case $C_7^{(5)}(3) = 2 + 3 + 4 + 3 = 12$. The $A_n(3)$'s are the column sums; for instance, $A_8(3) = 1 + 8 + 28 + 56 + 65 + 40 + 10 = 208$. If you toss a biased coin 8 times, we now have from (3) that the probability of obtaining no more than three consecutive heads is

$$1q^8 + 8pq^7 + 28p^2q^6 + 56p^3q^5 + 65p^4q^4 + 40p^5q^3 + 10p^6q^2.$$

The recursion for general $x < k < n$ is

$$C_n^{(k)}(x) = \sum_{j=0}^{x} C_{n-1-j}^{(k-j)}(x). \tag{4}$$

For $p \neq 1/2$, the form of $F_n'(x)$ (the probability that neither the longest run of heads nor the longest run of tails exceeds x) is more complicated but can be obtained from the same recursive idea. The result is omitted here; however when n is sufficiently large, the values that $F_n'(x)$ takes for $P(\text{heads}) = p$ are well approximated by the values realized for $F_n(x)$ for $P(\text{heads}) = \max(p, q)$. This is because when n is very large, the longest run will almost certainly be composed of whichever is more likely between heads and tails.

Properties of the Distribution of the Longest Run

A short computer program can easily be written using (1) or (4) that will rapidly generate the exact distribution of R_n for any moderate n. Perusal of the displays from such computations for $n = 1, 2, 3, \ldots$ reveals several unusual, even remarkable, features of longest run distributions. Detailed asymptotic analyses have clarified the precise nature of these attributes.

The log n law. Figure 3 shows the distribution of R_n for a fair coin for $n = 50$, 100 and 200. Immediately noticeable is that the distribution of the length of the longest run tends to shift towards larger values at a rate that is logarithmically related to n. A simple intuitive argument provides insight into why this phenomenon occurs.

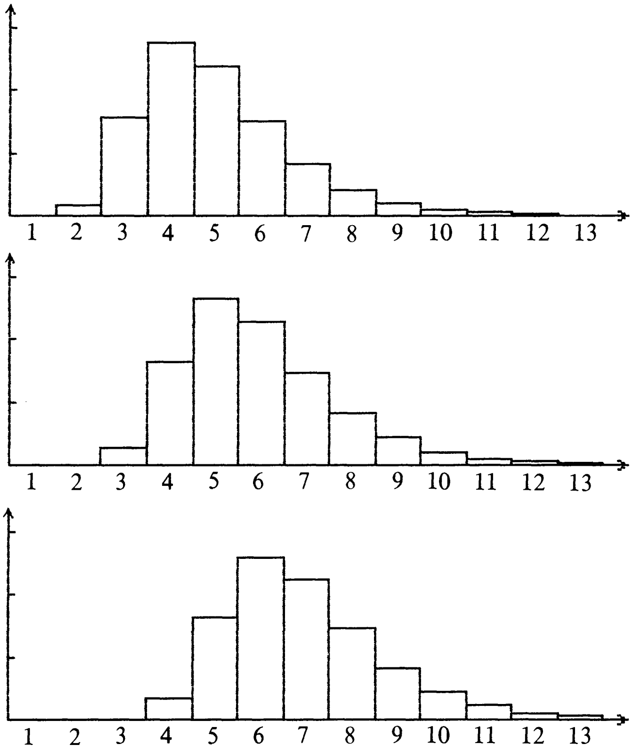


Figure 3
Distributions of R_n for (a) $n = 50$, (b) $n = 100$, (c) $n = 200$

Viewing each head run in a coin tossing sequence as the number of consecutive heads beginning with the first toss or immediately after tails occurs (allowing runs of length zero), there will be approximately nq head runs in all, since this is the expected number of tails. Around nqp of these head runs will contain at least one head, nqp^2 will be at least two heads long, and so forth. As long as nqp^x is greater than or equal to one, at least one run of length x or more can be expected; for larger values of x for which nqp^x falls below 1, obtaining a run as long as x is unlikely. Hence solving $nqp^{R_n} \approx 1$ for R_n gives a reasonable value for the typical length of the longest head run, namely

$$R_n \approx \log_{1/p}(nq).$$

For the case $p = 1/2$ we have $R_n \approx \log_2 n - 1$; this yields that R_n should be somewhere near 4.6, 5.6, 6.6 for $n = 50$, 100 and 200, respectively (compare to Figure 3). Rényi [**13**] proved the important result that for any given infinite sequence of tosses of a fair coin, the quantity $R_n/\log_2 n$ will converge to 1 with probability one. Numerous extensions have subsequently been treated; these include applications to Wiener and renewal processes and other stochastic processes as well as runs where the successive trials have more than two possible outcomes (see [**4**] and [**6**] for surveys), runs interrupted by a prescribed number of tails ([**8**], [**9**]), largest rectangles consisting entirely of 1's in a lattice (array) containing only 0's and 1's ([**7**], [**12**]), longest common subsequences contained in two sequences taking values in the same set ([**2**], [**3**]), and so forth. The extensiveness of the $\log n$ growth phenomenon for runs deserves much wider recognition.

Expectation and variance of the longest head run. One useful way to analyze head runs is by using *geometric* random variables to model the individual runs that comprise a sequence of coin tosses. The geometric, or waiting time, random variable is defined as the number of failures in a sequence of independent trials that occur before the first success.

In our situation we can identify heads with failure and tails with success. Since the approximately nq individual head runs (some possibly of length zero) in a sequence of tosses do not overlap, the longest head run can therefore be represented as essentially the maximum of nq independent geometric random variables. By using such a representation, the following asymptotic-based formulas for the expectation and variance of the longest head run can be derived:

$$ER_n = \log_{1/p}(nq) + \gamma/\ln(1/p) - 1/2 + r_1(n) + \varepsilon_1(n), \tag{5}$$

$$\operatorname{Var} R_n = \pi^2/6\ln^2(1/p) + 1/12 + r_2(n) + \varepsilon_2(n), \tag{6}$$

where $\gamma = 0.577\ldots$ is Euler's constant, $r_1(n)$ and $r_2(n)$ are very small (e.g., $|r_1(n)| \le 0.000016$, $|r_2(n)| \le 0.00006$ for all n when $p = 1/2$) periodic functions of $\log_{1/p} n$, and $\varepsilon_1(n)$ and $\varepsilon_2(n)$ tend to zero as $n \to \infty$. See **[8]** for details. These results were first obtained through the use of generating functions by Boyd **[5]** for the case $p = 1/2$; see also **[9]**.

Note that the leading term of ER_n is consistent with the heuristic argument given above for R_n. For $p = 1/2$ we get the simple approximation

$$ER_n \approx \log_2(n/2) + \gamma/\ln 2 - 1/2 \approx \log_2 n - 2/3.$$

Applying this to $n = 200$ we find that the expected length of the longest run of heads is approximately seven, while from (2) the expectation for the longest run of either pure heads or pure tails is about eight. Thus Sequence #1 given at the beginning is quite typical for real coin tossing experiments. Very few students who simulate 200 coin tosses list any runs longer than five.

The result for the variance (6) is quite remarkable for the property that it is essentially constant with respect to n. This means that we can predict the length of the longest run equally well by the $\log n$ formula, for, say, $n = 200$ as for $n = 10$ or $n = 2^{200}$! The next section explores this property further.

Prediction intervals. Just how accurately can we predict what the length of the longest run will be? Let us concentrate here on the case of a fair coin. From the asymptotic variance formula given above, the standard deviation of the longest run is approximately $(\operatorname{Var} R_n)^{1/2} \approx (\pi^2/6\ln^2 2 + 1/12)^{1/2} = 1.873$, an amazingly small value. This implies that the length of the longest run is quite predictable indeed; normally it is within about two of its expectation.

To further emphasize the predictability of the length of the longest run, we can also look at the probability that R_n will be contained within a small interval of possible values. Using asymptotic methods, the best interval of any given size can be found and the corresponding probability that R_n will fall within this interval can be computed. Calculations for finite values of n indicate that these asymptotic coverage probabilities are in fact slightly conservative. These probabilities are displayed in Table 1. The third line of the table, for example, shows that for any n it is possible to find an interval of length three for which the probability that R_n lies in this interval is at least 62.3%. Note that for every n more than 90% of the distributions of R_n and R'_n live on just six values, and 99% on ten values.

Table 1
Prediction Interval Probabilities for R_n ($p = 1/2$)

Width of interval	Minimum probability that R_n lies in the interval
1	23.6%
2	44.9%
3	62.3%
4	75.5%
5	84.6%
6	90.7%
7	94.5%
8	96.8%
9	98.2%
10	99.0%

A very easy rule of thumb is that the longest head run for a fair coin is very likely to be within three either way from the integer nearest to $\log_2(n/2)$. Applying this rule for $n = 200$, we find that reasonable limits for R_{200} are 4 and 10. The actual probability that the longest head run is between these values turns out to be 95.3%, which slightly exceeds the lower bound of 94.5% guaranteed by Table 1. For R'_{200}, the longest run of heads *or* of tails in 200 tosses, simply add one to each of the limits.

The question of a limiting distribution. Insights into the character of random phenomena are frequently obtained by looking at the asymptotic distributions of the random variables involved. The best known example of this is the central limit theorem, which says that under appropriate conditions, if the arithmetic mean of a collection of n random variables is standardized by subtracting its expectation and then dividing by the square root of its variance, then as n increases, the resulting quantity will have a distribution that approaches the standard normal distribution. The standardization is required in order to convert the mean to a new variable whose expectation and variance are stable as n increases to infinity.

Can we obtain, in similar fashion, a limiting distribution for the longest head run? The answer, strangely enough, is almost but not quite. Picture the probability histograms for $R_1, R_2, \dots$. (See Figure 3.) The expectation formula for R_n implies that these histograms drift steadily to the right at a rate governed by $\log_{1/p}(nq)$, while the near-constancy of the variance shows that they remain essentially stable in spread. We might hope, therefore, that the *aligned* probability histograms of $R_n - \log_{1/p}(nq)$ converge to a limiting histogram as $n \to \infty$. In fact, however, although the general *shape* of the aligned histograms stabilizes, the sequence contains a perpetual 'wobble' which cannot be eliminated even by additional manipulations. This phenomenon is explained a bit later.

The distribution to which the head run distributions are 'attempting' to converge is known as an *extreme value* distribution. This distribution arises under quite general conditions when the distribution of the maximum of a large number of independent random variables is studied. Since the length of the longest head run is the maximum of the lengths of the approximately nq component head runs, it should not be surprising that the extreme value distribution is involved here.

Specifically, consider a continuous random variable W_p whose cumulative distribution function is

$$F_{W_p}(x) = \exp^{-p^{x+1}}, \qquad -\infty < x < \infty.$$

The smooth curve in Figure 4 shows the probability density $f(x) = (d/dx)(F_{W_p}(x))$ of W_p for the case $p = 1/2$.

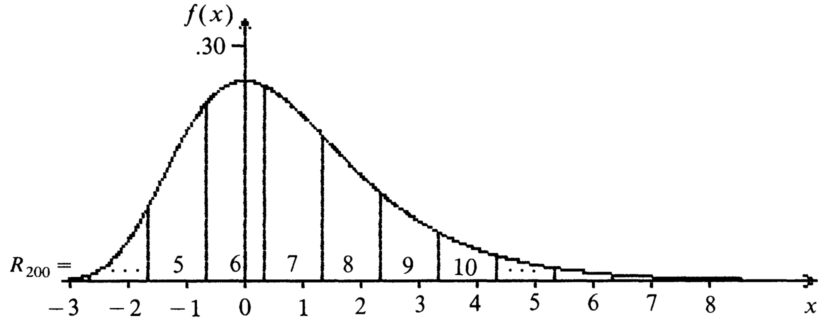


Figure 4
The approximating extreme value distribution for the longest run of heads in n tosses of a fair coin

A precise but rather complex description of the limiting behavior of R_n is furnished in [**8**]. The basic result is that the longest head run distribution satisfies the approximation

$$P(R_n = x) \approx P\left(x - \log_{1/p}(nq) < W_p \le x + 1 - \log_{1/p}(nq)\right) \qquad (7)$$

for $x = 0, 1, 2, \ldots$, with the error decreasing to zero as n tends to infinity.

Figure 4 illustrates how this works for the case of 200 tosses of a fair coin. The exact probabilities for R_n are asymptotically approximated by the areas contained between the vertical lines, which are located at the values of $x - \log_{1/p}(nq) = x - 6.644$ for $x = 0, 1, 2, \ldots$. Note that the mode of the extreme value density occurs at 0, which corresponds to the value $R_n = \log_{1/p}(nq)$, the conjectured 'typical' approximate length of the longest head run.

The skewness of the density shown in Figure 4 indicates that it is much more likely that the longest head run will be significantly longer than $\log_{1/p}(nq)$ than that it will be much shorter. To see why this is, note that there will be many times during a coin tossing sequence when a run of heads will begin to approach $\log_{1/p}(nq)$ (recall the heuristic argument given for the $\log n$ law), making it rather improbable that all such runs will fall much short, whereas if any ongoing run survives to length $\log_{1/p}(nq)$ it will not be that uncommon for a few more consecutive heads to occur before the run ends.

Notice that even Figure 1 resembles the extreme value curve to a reasonable degree. This shows that the distribution of R_n approaches its asymptotic form quite rapidly, i.e., even for very small values of n.

Table 2 shows exact values obtained from (1) and approximate values obtained from (7) for the distribution of R_{200} for a fair coin. Note that to obtain the approximate distribution of R_n for $p = 1/2$ and, say, $n = 200 \times 2^{10} = 204{,}800$, all one has to do is to add 10 to each value in the x column of Table 2.

Table 2
Exact and Approximate Probabilities for $R_{200}(p=1/2)$

x	$P(R_{200}=x)$ (Exact)	$P(R_{200}=x)$ (Approx.)
0–3	.001	.002
4	.033	.042
5	.165	.166
6	.257	.248
7	.224	.219
8	.146	.146
9	.083	.084
10	.044	.045
11	.023	.024
12	.011	.012
> 12	.012	.012

Now let us return to the problem of the lack of a proper limiting distribution. Although the extreme value distribution provides an asymptotic approximation, the sequence of R_n distributions retains a 'wobble' whose frequency decreases geometrically but which persists forever with constant amplitude. The reason for this phenomenon is that while the R_n distributions want to shift smoothly to the right according to the rate $\log_{1/p}(nq)$, they are constrained to live on the integers. The dividing lines illustrated in Figure 4 are constantly shifting to the left as n increases, and are aligned for two values of n only if the ratio of these values is a power of $1/p$.

Hence the most that can be said is that there are *subsequences* $\{n_i,\ i=1,2,\dots\}$ for which the R_{n_i} distributions possess a limit. An example for $p=1/2$ is the subsequence $n_i=2^i$, $i=1,2,\dots$. Figure 5 shows the outlines of the aligned probability histograms of $R_n-\log_2(n/2)$ for two members of this subsequence along with the histogram for an intermediate value of n not belonging to the subsequence. Note the close agreement between the histograms for $n=32$ and $n=64$; for larger values of $n=2^i$, the histograms are indistinguishable.

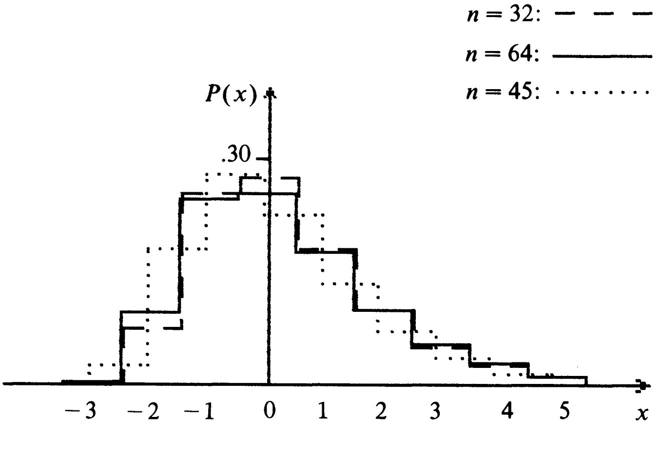


Figure 5
Probability histograms for $R_n-\log_2(n/2)$ for a fair coin

A computer animation program showing the sequence of head run distributions for a fair coin (written in BASIC for IBM and compatible PC's) is available from the author. It displays the exact probability histograms for R_n for $n = 1, 2, \ldots, 100$, obtained from (1) and corrected for the $\log n$ drift to the right, with the corresponding extreme value density overlaid.

Extensions to other runs-related phenomena; applications. Many other types of runs-related phenomena have been studied. For example, Gordon, Schilling and Waterman [**8**] give results not only for the case of pure head runs but also for head runs interrupted by a specified number of tails. Other recent work includes further analysis of runs that may or may not overlap themselves [**9**], largest cubes of ones in a d-dimensional random lattice of zeros and ones ([**7**], [**12**]), and longest common subsequences in two strings defined over an alphabet of l letters, allowing shifts ([**2**], [**3**]). The latter situation has important applications in molecular biology to the matching of DNA sequences—for example, from corresponding genetic sites in related species—and illustrates some of the general structure that tends to occur in such problems. Roughly speaking, the corresponding form of the $\log n$ law can often be guessed by reasoning analogous to that given for the $\log n$ growth rule for head runs, that is, by treating occurrences of specific patterns at different locations as if they were probabilistically independent. (Though they are often *not* independent because of possible overlaps, it can be shown quite generally that at least asymptotically the dependence is negligible.)

In the case of comparing two DNA strings of lengths m and n, there are mn opportunities for a nucleotide of the first sequence to match a nucleotide of the second, i.e., mn opportunities to start a 'run' (common subsequence). Thus a plausible value for the length of the longest common subsequence, allowing shifts when aligning the two sequences, is $\log_{1/p}(mnq)$, where p is now the probability of a *match* between the nucleotides at two sites selected arbitrarily, one from each sequence, and $q = 1 - p$ as before. This indeed turns out to give the correct $\log n$ law for sequence matching.

Numerous other variations and extensions can be considered including first-order Markov chains; that is, sequences in which each outcome is dependent on the previous one (see [**15**] and [**17**]) and situations in which the probability of a given outcome is not the same for all trials [**19**]. In particular, the winning and losing streaks of a team or individual in some sport might be modeled this way, with the probability of winning determined both by the location of the sporting event ('home' or 'road') and/or the strengths of the various opponents.

It is also interesting to note the relationship of the results for the longest head run to the following coin-tossing game, which was studied by Kinney [**10**]: Toss N coins simultaneously; then toss again only those that come up tails the first time. Continuing in this manner until each of the coins shows heads, Kinney wondered how many stages of the process would be required.

Once again we are looking at the distribution of the maximum of several geometric random variables, as in the case of the longest head run. The only difference is that here the number of these variables is fixed at N, whereas for the longest run this number was random, being determined by the number of tails that occur in n tosses of a single coin. Thus we should expect very similar results to those given above with N replacing nq in each of the formulas. This indeed turns out to be the case (see [**16**]).

The variety of potential applications of runs theory is virtually boundless. Some of the more intriguing include handwriting analysis by means of digitized scanning

[1], hydrologic runs (floods and droughts; see for example [18]), and studies of the pattern of capture of prey species [11].

Acknowledgment. This work was supported in part by the California State University, Northridge.

References

1. B. Arazi, Handwriting identification by means of run-length measurements, *IEEE Transactions on Systems, Man and Cybernetics* 7(12) (1977) 878–881.
2. R. Arratia, L. Gordon, and M. S. Waterman, An extreme value theory for sequence matching, *Annals of Statistics* 14 (1986) 971–993.
3. R. Arratia and M. S. Waterman, An Erdös-Rényi law with shifts, *Advances in Mathematics* 55 (1985) 13–23.
4. S. A. Book, Large deviation probabilities and the Erdös-Rényi law of large numbers, *Canadian Journal of Statistics* 4 (1976) 185–209.
5. D. W. Boyd, Losing runs in Bernoulli trials, unpublished manuscript (1972).
6. S. Csörgo, Erdös-Rényi laws, *Annals of Statistics* 7(4) (1979) 772–787.
7. R. Darling and M. S. Waterman, Extreme value distribution for the largest cube in a random lattice, *SIAM Journal of Applied Mathematics* 46 (1986) 118–132.
8. L. Gordon, M. F. Schilling, and M. S. Waterman, An extreme value theory for long head runs, *Probability Theory and Related Fields* 72 (1986) 279–287.
9. L. J. Guibas and A. M. Odlyzko, Long repetitive patterns in random sequences, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebeite* 53 (1980) 241–262.
10. J. Kinney, Tossing coins until all are heads, *Mathematics Magazine* 51(3) (1978) 184–186.
11. J. N. McNair, A stochastic foraging model with predator training effects: I. Functional response, switching and run lengths, *Theoretical Population Biology* 17(2) (1980) 141–166.
12. T. Nemetz and N. Kusolitsch, On the longest run of coincidences, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebeite* 61 (1982) 59–73.
13. A. Rényi, *Probability Theory*, Akad. Kiadó, Budapest, 1970.
14. P. Révész, Strong theorems on coin tossing, *Proceedings of the International Congress of Mathematicians*, Helsinki (1978) 749–754.
15. S. S. Samarova, On the asymptotic behaviour of the maximal sojourn time of an ergodic Markov chain in a fixed state, *Russian Math Surveys* 35(6) (1980) 103–104.
16. E. F. Schuster, On overwhelming numerical evidence in the settling of Kinney's waiting time conjecture, *SIAM Journal of Statistical Computing*, 6(4) (1985) 977–982.
17. S. J. Schwager, Run probabilities in sequences of Markov-dependent trials, *Journal of the American Statistical Association* 78 (1983) 168–175.
18. Z. Sen, Statistical analysis of hydrologic critical droughts, *Journal of the Hydraulics Division* 106(HY1) (1980) 99–115.
19. Y. L. Tong, A rearrangement inequality for the longest run with an application to network reliability, *Journal of Applied Probability* 22 (1985) 386–393.

On the Edge at .05

Frederick Mosteller tells us that if you toss a coin repeatedly in a college class and after each toss ask the class if there is anything suspicious going on, "hands suddenly go up all over the room" after the fifth head or tail in a row. There happens to be only 1 chance in 16—.0625, not far from .05, or 5 chances in 100–that five heads or tails in a row will show up in five tosses, "so there is some empirical evidence that the rarity of events in the neighborhood of .05 begins to set people's teeth on edge."

Victor Cohn, *News and Numbers*, Iowa State University Press, 1989