

Examining Probabilistic Runs

NCSSM Teaching Contemporary Mathematics Conference
January 27-28, 2006

Christopher H. Jones
Horace Mann School
New York, New York

Table of Contents

I Introduction	p. 2
II The Run of Heads Problem	pp. 3-10
III Generalizations/Extensions	pp. 11-12
IV Solutions to Generalizations/Extensions	pp. 13-17
V A Couple of Applications	pp. 18-22

Christopher H. Jones
(w) 718-432-3938
chris_jones@horacemann.org

Introduction

For the past couple of summers, I have taught a "Problem-Solving" course to accelerated middle schoolers at Phillips Exeter Academy in New Hampshire. The kids take three loosely "linked" courses (one in math, one in computer science, and one in physics/robotics). In the computer class, the students do a unit on programming simple experiment simulations with the goal of approximating specific probabilities via a Monte Carlo method.

One of the experiments the students simulate is flipping a fair coin 1000 times. They are asked to approximate the probability that within the 1000 flips there exists a string of at least 10 consecutive heads. Quite naturally, they run the experiment thousands of time on a loop (not so hard to do with a fast computer and reasonably efficient code) and keep track of the ratio of successes to total number of trials. When they are done, they have their estimate.

Several of the students (and the computer science instructor) inquired whether I knew how to compute this probability. I had thought a little bit about the well-known "consecutive runs" problems in the past, but had given up examining them too closely, believing the counting techniques involved would be too difficult. But this time, given the enthusiasm of my students to learn more about this type of problem, I was motivated to learn a good deal more about such problems.

In this handout, I have tried to give the reader the tools to examine this problem and related ones. At the end I have included two applications: the first having to do with digit runs in the expansion of π ; the second having to do with making a prediction about an event in Barry Bonds' 2004 MVP season. Professional sports statisticians (and fans) are obsessed with streaks, and I believe there is some fun to be had in using the math in this handout to examine the likelihoods of these phenomena.

The Run of Heads Problem

Okay, let's do some math!

We begin by devising some notation that will allow for greater ease in manipulating the ideas involved. We define the function

$H(r, n)$ = the probability that a run of least r consecutive heads appears in a string of n flips of a fair coin,

where, for obvious reasons, $r, n \in \mathbb{Z}$ and $r, n \geq 0$.

Using this new notation, the problem-at-hand is to compute the value of

$$H(10, 1000).$$

If one attacks this problem head-on, a natural first step is to consider

$$H(10, 1000) = \frac{\text{Number of 1000-letter strings of } T\text{'s and } H\text{'s that contain a string of at least 10 consec } H\text{'s}}{\text{Number of 1000-letter strings containing only } T\text{'s and } H\text{'s}} = \frac{\text{Number of 1000-letter strings of } T\text{'s and } H\text{'s that contain a string of at least 10 consec } H\text{'s}}{2^{1000}}$$

The problem is now purely combinatorial. To evaluate the desired probability we need to count the number of 1000-letter strings made exclusively of T 's and H 's that contain a string of at least 10 H 's. This is no trivial matter! Think about the number of qualitatively different ways in which this could occur:

(i) the string contains exactly one run of at least 10 consec H 's (but is it 10 in total or is it embedded in a longer string? – that's more sub-cases to consider).

(ii) the string contains exactly two separated strings of at least 10 consec H 's (but are they embedded in longer strings? – that's many more sub-cases to consider).

⋮

Very quickly, one's head begins to hurt and the realization sets in that this is not the way to go – particularly if there is the desire to generalize one's results. So, what to do?

Recursion to the rescue!

Rather than examining the combinatorial problem associated with the numerator on the last page, we will take a close look at the probabilities [the $H(r,n)$ function] with an eye to recursion.

Consider the following clever "splitting" of the problem¹:

$$H(10,n) = \overbrace{\left(\begin{array}{l} \text{The probability that there} \\ \text{is a run of at least 10 consec H's} \\ \text{in the } n \text{ flips AND the last flip} \\ \text{is NOT "pivotal" in its creation.} \end{array} \right)}^{\text{Last flip NOT "pivotal"}} + \overbrace{\left(\begin{array}{l} \text{The probability that there is a} \\ \text{run of at least 10 consec H's} \\ \text{in the } n \text{ flips AND the last} \\ \text{flip is "pivotal" in its creation} \end{array} \right)}^{\text{Last flip "pivotal"}}.$$

We examine the last flip NOT "pivotal" case first, as it is easier to dissect:

$$\begin{aligned} \text{Last flip NOT pivotal} &= \left(\begin{array}{l} \text{The probability that there} \\ \text{is a run of at least 10 consec H's} \\ \text{in the } n \text{ flips AND the last flip} \\ \text{is NOT "pivotal" in its creation.} \end{array} \right) \\ &= \left(\begin{array}{l} \text{The prob that there is a} \\ \text{run of at least 10 consec H's} \\ \text{in the first } n-1 \text{ flips} \\ \text{AND} \\ \text{the } n\text{th flip is either H or T} \end{array} \right) \\ &= H(10,n-1) \cdot (\text{prob the } n\text{th flip is either H or T}) \\ &= H(10,n-1) \cdot (1) \\ &= H(10,n-1) \end{aligned}$$

¹ I found this idea in J.V. Uspensky's 1937 edition of "Introduction to Probability."

Now we tackle the last flip is "pivotal" case:

$$\begin{aligned}
 \text{Last flip pivotal} &= \left(\begin{array}{l} \text{The probability that there is a} \\ \text{run of at least 10 consec H's} \\ \text{in the } n \text{ flips} \\ \text{AND} \\ \text{the last flip is "pivotal" in its creation} \end{array} \right) \\
 &= \left(\begin{array}{l} \text{The probability that the one and only string} \\ \text{of 10 consec H's occurs in the final 10 flips} \end{array} \right) \\
 &= \left(\begin{array}{l} \text{The probability that the } n \text{ flips} \\ \text{look like the string pictured below } \downarrow \end{array} \right)
 \end{aligned}$$

These n-11 flips do NOT contain a run of at least 10 H's

The final flip saves the day and creates the first and only run of 10 H's

 ... T H H H H H H H H H H

$$\begin{aligned}
 &= \left(\begin{array}{l} \text{The prob that there is NOT} \\ \text{a run of at least 10 consec} \\ \text{H's in the first } n-11 \text{ flips} \end{array} \right) \cdot \left(\begin{array}{l} \text{The prob that} \\ \text{flip } n \text{ is H and} \\ \text{flip } n-1 \text{ is H and} \\ \vdots \\ \text{flip } n-9 \text{ is H and} \\ \text{flip } n-10 \text{ is T} \end{array} \right) \\
 &= \left[1 - \left(\begin{array}{l} \text{The prob that there IS} \\ \text{a run of at least 10 consec} \\ \text{H's in the first } n-11 \text{ flips} \end{array} \right) \right] \cdot \left(\begin{array}{l} \text{The prob that} \\ \text{flip } n \text{ is H and} \\ \text{flip } n-1 \text{ is H and} \\ \vdots \\ \text{flip } n-9 \text{ is H and} \\ \text{flip } n-10 \text{ is T} \end{array} \right) \\
 &= [1 - H(10, n-11)] \cdot \left(\frac{1}{2} \right)^{11}
 \end{aligned}$$

Putting the two cases together, we arrive at

$$H(10,n) = \overbrace{H(10,n-1)}^{\text{prob that last flip is NOT pivotal}} + \overbrace{[1 - H(10,n-1)] \cdot \left(\frac{1}{2}\right)^{11}}^{\text{prob that last flip is pivotal}}$$

This is a straightforward recurrence relation of degree 11. If we can determine the 11 required initial conditions below

$$\underbrace{\{H(10,0), H(10,1), \dots, H(10,10)\}}_{\text{first 11 values of the } H(10,n) \text{ function}},$$

then we can get this recursion started. Luckily, all 11 are trivially easy to compute as

$H(r,n) = 0$ when $r > n$ and $H(r,n) = \left(\frac{1}{2}\right)^r$ when $r = n$ (see table below).

n	$H(10,n)$
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	$(1/2)^{10}$

Now the recursion kicks in.

$$\begin{aligned}
 H(10,n) &= H(10,n-1) + [1 - H(10,n-1)] \cdot \left(\frac{1}{2}\right)^{11} \\
 \rightarrow H(10,11) &= H(10,10) + [1 - H(10,0)] \cdot \left(\frac{1}{2}\right)^{11} \\
 &= \left(\frac{1}{2}\right)^{10} + [1 - 0] \left(\frac{1}{2}\right)^{11} = \frac{3}{2^{11}}
 \end{aligned}$$

$$\begin{aligned} \rightarrow H(10,12) &= H(10,11) + [1 - H(10,1)] \cdot \left(\frac{1}{2}\right)^{11} \\ &= \frac{3}{2^{11}} + [1 - 0] \left(\frac{1}{2}\right)^{11} = \frac{4}{2^{11}} = \frac{8}{2^{12}} \\ &\vdots \end{aligned}$$

Enough calculating by hand, it is time to let a machine take over and do the heavy computational lifting for us.

Excel is by nature recursion-friendly! To begin, we set up the 11 initial conditions in the first 11 cells (those corresponding to n values of 0-10). Next we need to translate our recurrence relation into Excel-ese:

$$\text{Contents of cell} = \left(\begin{array}{c} \text{contents of cell} \\ \text{one above} \end{array} \right) + \left(1 - \left(\begin{array}{c} \text{contents of cell} \\ \text{eleven above} \end{array} \right) \right) \cdot \left(\frac{1}{2}\right)^{11}.$$

This is accomplished by typing into cell B13 the formula you see below:

	A	B
1	n	$H(10,n)$
2	0	0
3	1	0
4	2	0
5	3	0
6	4	0
7	5	0
8	6	0
9	7	0
10	8	0
11	9	0
12	10	$=(1/2)^{10}$
13	11	$= B12 + (1 - B2) * (1/2)^{11}$

To compute values deeper into the sequence, we highlight the cells A13 and B13, and then do a "formula drag" as far down as we desire. (see the extended table on the next page).

	A	B
1	n	$H(10,n)$
2	0	0
3	1	0
4	2	0
5	3	0
6	4	0
7	5	0
8	6	0
9	7	0
10	8	0
11	9	0
12	10	0.00097656
13	11	0.00146484
14	12	0.00195313
15	13	0.00244141

⋮

992	990	0.38242610
993	991	0.38272913
994	992	0.38303202
995	993	0.38333475
996	994	0.38363734
997	995	0.38393978
998	996	0.38424207
999	997	0.38454421
1000	998	0.38484621
1001	999	0.38514805
1002	1000	0.38544975

We've done it (or, rather, Excel has done it). We have the value we sought:

$$H(10,1000) \approx 0.3854 .$$

Can we use the TI-83/84 to do the same type of recursion? We could attempt to use SEQ mode and give it a shot, but this feature does not accept recurrence relations of degree higher than 2, and we are dealing with degree 11. It makes more sense to write brief programs to do the recursion.

Consider the following two programs:

```

Program:AAHEADRN
(1) :ClrHome
(2) :Input "RUN LENGTH:",R
(3) :Input "NUM FLIPS:",N
(4) :R→dim(L1)
(5) :Fill(0,L1)
(6) :(1/2)R→L1(R+1)
(7) :
(8) :For(A,R+2,N+1,1)
(9) :L1(A-1)+(1-L1(A-11))*(1/2)11→L1(A)
(10) :End
(11) :
(12) :Disp ""
(13) :Disp "PROB OF RUN IS:"
(14) :Disp L1(N+1)

```

```

Program:AAHDRN2
(1) :ClrHome
(2) :Input "RUN LENGTH:",R
(3) :Input "NUM FLIPS:",N
(4) :R→dim(L1)
(5) :Fill(0,L1)
(6) :(1/2)R→L1(R+1)
(7) :
(8) :For(A,R+2,N+1,1)
(9) :L1(R+1)+(1-L1(1))*(1/2)(R+1)→L1(R+2)
(10) :
(11) :For(B,1,R+1)
(12) :L1(B+1)→L1(B)
(13) :End
(14) :
(15) :End
(16) :
(17) :Disp ""
(18) :Disp "PROB OF RUN IS:"
(19) :Disp L1(R+2)

```

Both programs compute $H(R,N)$ with the following differences:

- AAHEADRN stores all of the intermediate values $H(R, 0), \dots, H(R, N)$ in L_1 while AAHDRN does not.
- AAHEADRN will not work for values of $N \geq 999$ as the TI-83/84 calculators have a maximum list size of 999 elements.
- AAHDRN2 will work (rather slowly) for values of $N \geq 999$. It requires a list size of only $R+2$ elements. This program computes the next term in the $H(R,N)$ sequence, and then recalibrates the list of relevant elements (this takes place in lines (11) – (13)) so that a minimum of memory space is required. The cost, of course, is computing time, as this recalibration creates many extra steps in the algorithm. It took my TI-84 Plus approximately 4 minutes to compute the correct value of $H(10,1000)$.

While writing such programs is definitely a worthwhile exercise, I think that the ease with which Excel handles this problem makes it the preferable technological tool for this area of study.

Generalizations/Extensions

The problem on which I have focused can be extended and generalized in several ways.

First, we should generalize the formula for $H(10, n)$ to $H(r, n)$. This is a straightforward matter of replacing all instances of 10 in our $H(10, n)$ formula with r .

This yields a sequence with initial conditions

$$\begin{aligned} H(r, 0) &= 0 \\ H(r, 1) &= 0 \\ &\vdots \\ H(r, r-1) &= 0 \\ H(r, r) &= \left(\frac{1}{2}\right)^r \end{aligned}$$

and the recurrence relation that holds for $n > r$ is

$$H(r, n) = H(r, n-1) + [1 - H(r, n-r-1)] \cdot \left(\frac{1}{2}\right)^{r+1}$$

Consider the following extensions:

1.
 - A) In n flips of a fair coin, what is the probability that the maximum run of heads will be of length r ?
 - B) In n flips of a fair coin, what is the expected value of the maximum run of heads?

2.
 - A) In n flips of a fair coin, what is the probability there will be a run of at least r consecutive heads or tails?
 - B) In n flips of a fair coin, what is the probability that the maximum run of heads or tails will be of length r ?
 - C) In n flips of a fair coin, what is the expected value of the maximum run of heads or tails?

Consider the following generalizations:

3.
 - A) In n identical trials of an experiment (**each with probability of success = q**) what is the probability of a run of successes of at least length r ?
 - B) In n identical trials of an experiment (each with probability of success = q) what is the probability the maximum run of successes will be of length r ?
 - C) In n identical trials of an experiment (each with probability of success = q) what is the expected value of the maximum run of successes?

4.
 - A) In n identical trials of an experiment (each with probability of success = q) what is the probability of a run of successes or failures of at least length r ?
 - B) In n identical trials of an experiment (each with probability of success = q) what is the probability the maximum run of successes or failures will be of length r ?
 - C) In n identical trials of an experiment (each with probability of success = q) what is the expected value of the maximum run of successes or failures?

More general still:

5.
 - A) In n identical trials of an experiment (**with c equally likely outcomes**) what is the probability of a run of any type of outcome of at least length r ?
 - B) In n identical trials of an experiment (with c equally likely outcomes) what is the probability that the maximum run of any type of outcome will be of length r ?
 - C) In n identical trials of an experiment (with c equally likely outcomes) what is the expected value of the length of the maximum run of any type of outcome?

Solutions to Generalizations/Extensions

1. A) We define $MaxRunH(r, n) =$ the prob that in n flips of a fair coin the maximum run of consec heads has length r . We arrive at

$$\boxed{MaxRunH(r, n) = H(r, n) - H(r + 1, n)}$$

1. B) We define $ExpMaxRunH(n) =$ the expected value of the length of the maximum run of heads in n flips of a fair coin.

$$\begin{aligned} ExpMaxRunH(n) &= \sum_{i=1}^n i \cdot MaxRunH(i, n) \\ &= \sum_{i=1}^n i \cdot [H(i, n) - H(i + 1, n)] \\ &= \sum_{i=1}^n i \cdot H(i, n) - \sum_{i=1}^n i \cdot H(i + 1, n) \\ &= \sum_{i=1}^n i \cdot H(i, n) - \sum_{i=2}^{n+1} (i - 1) \cdot H(i, n) \\ &= 1 \cdot H(1, n) + \sum_{i=2}^n i \cdot H(i, n) - \sum_{i=2}^n (i - 1) \cdot H(i, n) + n \cdot \overbrace{H(n + 1, n)}^{This=0} \\ &= H(1, n) + \sum_{i=2}^n [i \cdot H(i, n) - (i - 1) \cdot H(i, n)] \\ &= H(1, n) + \sum_{i=2}^n H(i, n) \\ &= \sum_{i=1}^n H(i, n) \end{aligned}$$

So, we have derived the formula

$$\boxed{ExpMaxRunH(n) = \sum_{i=1}^n H(i, n)}$$

2. A) We define $HorT(r, n) =$ the prob that n flips of a fair coin contains a run of at least r consecutive heads or tails.

The initial conditions are:

$$\begin{aligned} HorT(r, 0) &= 0 \\ HorT(r, 1) &= 0 \\ &\vdots \\ HorT(r, r-1) &= 0 \\ HorT(r, r) &= \frac{2}{2^r} \end{aligned}$$

and the recurrence relation that applies for $n > r$ is

$$HorT(r, n) = HorT(r, n-1) + [1 - HorT(r, n-r)] \cdot \left(\frac{1}{2}\right)^r$$

Notice that this recurrence relation is of degree one smaller than that used to generate $H(r, n)$.

2. B) We define $MaxRunHorT(r, n) =$ the prob that in n flips of a fair coin the maximum run of heads or tails has a length of r .

$$MaxRunHorT(r, n) = HorT(r, n) - HorT(r+1, n)$$

2. C) We define $ExpMaxRunHorT(n) =$ the expected value of the length of the maximum run of heads or tails in n flips of a fair coin. In a fashion parallel to that used to derive the solution to #1A, we get

$$ExpMaxRunHorT(n) = \sum_{i=1}^n HorT(i, n)$$

3. A) We define $ProbRun(r, q, n)$ = the probability that there will be a run of at least r consecutive successes (**each with probability = q**) in n identical trials of an experiment.

We have the initial conditions

$$\begin{aligned} ProbRun(r, q, 0) &= 0 \\ ProbRun(r, q, 1) &= 0 \\ &\vdots \\ ProbRun(r, q, r-1) &= 0 \\ ProbRun(r, q, r) &= q^r \end{aligned}$$

and the recurrence relation that holds for $n > r$ is

$$ProbRun(r, q, n) = ProbRun(r, q, n-1) + [1 - ProbRun(r, q, n-r-1)] \cdot (1-q) \cdot q^r$$

3. B) We define $ProbMaxRun(r, q, n)$ = the prob that the max run of successes (each with prob = q) in n identical trials of an experiment will have length r .

$$ProbMaxRun(r, q, n) = ProbRun(r, q, n) - ProbRun(r+1, q, n)$$

3. C) We define $ExpMaxRun(q, n)$ = the expected value of the length of the max run of successes in n identical trials of an experiment (each with prob of success = q).

$$ExpMaxRun(q, n) = \sum_{i=1}^n ProbRun(i, q, n)$$

4. A) We define $ProbRunEith(r, q, n) =$ the prob that in n identical trials of an experiment (each with prob of success = q) there is a run of either successes or failures of at least length r .

We have the initial conditions:

$$\begin{aligned} ProbRunEith(r, q, 0) &= 0 \\ ProbRunEith(r, q, 1) &= 0 \\ &\vdots \\ ProbRunEith(r, q, r-1) &= 0 \\ ProbRunEith(r, q, r) &= q^r + (1-q)^r \end{aligned}$$

and the recurrence relation that holds for $n > r$ is

$$ProbRunEith(r, q, n) = ProbRunEith(r, q, n-1) + [1 - ProbRunEith(r, q, n-r)] \cdot [q(1-q)^r + (1-q)q^r]$$

4. B) We define $ProbMaxRunEith(r, q, n) =$ the prob that in n identical trials of an experiment (each with prob of success = q) there is a max run of either successes or failures of at least length r .

$$ProbMaxRunEith(r, q, n) = ProbRunEith(r, q, n) - ProbRunEith(r+1, q, n)$$

4. C) We define $ExpMaxRunEith(q, n) =$ the expected value of the max run of either successes or failures in n identical trials of an experiment (each with prob of success = q).

$$ExpMaxRunEith(q, n) = \sum_{i=1}^n ProbRunEith(i, q, n)$$

5. A) We define $ProbRunAny(r, c, n)$ = the prob that in n trials of an experiment (with c equally likely outcomes) there is a run of any type of outcome of at least length r .

We have the initial conditions

$$\begin{aligned} ProbRunAny(r, c, 0) &= 0 \\ ProbRunAny(r, c, 1) &= 0 \\ &\vdots \\ ProbRunAny(r, c, r-1) &= 0 \\ ProbRunAny(r, c, r) &= \left(\frac{c}{c}\right) \cdot \left(\frac{1}{c}\right)^{r-1} \end{aligned}$$

and the recurrence relation that holds for $n > r$ is

$$ProbRunAny(r, c, n) = ProbRunAny(r, c, n-1) + \left[1 - ProbRunAny(r, c, n-r)\right] \cdot \left(\frac{c-1}{c}\right) \cdot \left(\frac{1}{c}\right)^{r-1}$$

5. B) We define $ProbMaxRunAny(r, c, n)$ = the prob that in n identical trials of an experiment (with c equally likely outcomes) the max run of any type of outcome has length r .

$$ProbMaxRunAny(r, c, n) = ProbRunAny(r, c, n) - ProbRunAny(r+1, c, n)$$

5. C) We define $ExpMaxRunAny(c, n)$ = the expected value of the max run of any type of outcome in n identical trials of an experiment (with c equally likely outcomes).

$$ExpMaxRunAny(c, n) = \sum_{i=1}^n ProbRunAny(i, c, n)$$

A Couple of Applications

1. Examining π

If one examines a string of randomly generated integers (within a certain bound), there are probabilities associated with whether or not that string contains a "digit run" of at least a particular size.

I thought it would be interesting to put π to the test. I decided to examine chunks of equal size in the decimal expansion of π . I wanted to compute the frequency with which digit runs did or did not occur in these chunks and then compare it to the probabilities I generated using the recursive formula I in #5A.

I used *Mathematica* to print out several pages of digits of π . I noticed that the particular default formatting on my computer led to 85 digits of π per line of printout. Then, I went digit run hunting. I decided to look for digit runs of at least length 3. A line that contained such a run got a check (otherwise not). If π were to exhibit random behavior, then the probability that an individual line in my printout got a check should be

$$ProbRunAny(3,10,85).$$

This represents the probability that in 85 trials of an identical experiment (with 10 equally likely outcomes) there will be a run of any kind of at least length 3.

Using Excel and following the template set up in our computation of $H(10,1000)$ it is a straightforward matter to generate

$$ProbRunAny(3,10,85) \approx 0.53474$$

Okay, we have a theoretical probability. How does π stack up to this probability? On the next page, you see an actual π printout. In these first 55 lines, there are 28 "hits" and 27 "misses," leading to a ratio of

$$\frac{28}{55} \approx 0.5091$$

3.141592653589793238462643383279502884197169399375105820974944592307816406286208998628
0348253421170679821480865132823066470938446095505822317253594081284811174502841027019
3852110555964462294895493038196442881097566593344612847564823378678316527120190914564
8566923460348610454326648213393607260249141273724587006606315588174881520920962829254
0917153643678925903600113305305488204665213841469519415116094330572703657595919530921
8611738193261179310511854807446237996274956735188575272489122793818301194912983367336
2440656643086021394946395224737190702179860943702770539217176293176752384674818467669
4051320005681271452635608277857713427577896091736371787214684409012249534301465495853
7105079227968925892354201995611212902196086403441815981362977477130996051870721134999
9998372978049951059731732816096318595024459455346908302642522308253344685035261931188
1710100031378387528865875332083814206171776691473035982534904287554687311595628638823
5378759375195778185778053217122680661300192787661119590921642019893809525720106548586
3278865936153381827968230301952035301852968995773622599413891249721775283479131515574
8572424541506959508295331168617278558890750983817546374649393192550604009277016711390
0984882401285836160356370766010471018194295559619894676783744944825537977472684710404
7534646208046684259069491293313677028989152104752162056966024058038150193511253382430
0355876402474964732639141992726042699227967823547816360093417216412199245863150302861
8297455570674983850549458858692699569092721079750930295532116534498720275596023648066
5499119881834797753566369807426542527862551818417574672890977772793800081647060016145
2491921732172147723501414419735685481613611573525521334757418494684385233239073941433
3454776241686251898356948556209921922218427255025425688767179049460165346680498862723
2791786085784383827967976681454100953883786360950680064225125205117392984896084128488
6269456042419652850222106611863067442786220391949450471237137869609563643719172874677
6465757396241389086583264599581339047802759009946576407895126946839835259570982582262
0522489407726719478268482601476990902640136394437455305068203496252451749399651431429
8091906592509372216964615157098583874105978859597729754989301617539284681382686838689
4277415599185592524595395943104997252468084598727364469584865383673622262609912460805
1243884390451244136549762780797715691435997700129616089441694868555848406353422072225
8284886481584560285060168427394522674676788952521385225499546667278239864565961163548
8623057745649803559363456817432411251507606947945109659609402522887971089314566913686
7228748940560101503308617928680920874760917824938589009714909675985261365549781893129
7848216829989487226588048575640142704775551323796414515237462343645428584447952658678
2105114135473573952311342716610213596953623144295248493718711014576540359027993440374
2007310578539062198387447808478489683321445713868751943506430218453191048481005370614
6806749192781911979399520614196634287544406437451237181921799983910159195618146751426
9123974894090718649423196156794520809514655022523160388193014209376213785595663893778
7083039069792077346722182562599661501421503068038447734549202605414665925201497442850
7325186660021324340881907104863317346496514539057962685610055081066587969981635747363
8405257145910289706414011097120628043903975951567715770042033786993600723055876317635
9421873125147120532928191826186125867321579198414848829164470609575270695722091756711
6722910981690915280173506712748583222871835209353965725121083579151369882091444210067
5103346711031412671113699086585163983150197016515116851714376576183515565088490998985
9982387345528331635507647918535893226185489632132933089857064204675259070915481416549
859461637180270981994309924488957571282890592323260972997120844335732654893823911932
5974636673058360414281388303203824903758985243744170291327656180937734440307074692112
0191302033038019762110110044929321516084244485963766983895228684783123552658213144957
6857262433441893039686426243410773226978028073189154411010446823252716201052652272111
6603966655730925471105578537634668206531098965269186205647693125705863566201855810072
9360659876486117910453348850346113657686753249441668039626579787718556084552965412665
4085306143444318586769751456614068007002378776591344017127494704205622305389945613140
7112700040785473326993908145466464588079727082668306343285878569830523580893306575740
6795457163775254202114955761581400250126228594130216471550979259230990796547376125517
6567513575178296664547791745011299614890304639947132962107340437518957359614589019389
7131117904297828564750320319869151402870808599048010941214722131794764777262241425485
4540332157185306142288137585043063321751829798662237172159160771669254748738986654949

Not bad, but I wanted to look deeper. After several more pages worth of run hunting, I concluded (with very bleary eyes) that in the first 165 lines, there are 80 "hits" and 85 "misses," leading to a ratio of

$$\frac{80}{165} \approx 0.4848$$

Hmm... what to think? I was tired of looking at digits, so I thought it would make more sense to write a *Mathematica* Program to look at, say, the first 1000 strings of 85 digits of π . After much fumbling about, I succeeded in writing a very inefficient program that took even the speedy *Mathematica* a good deal of time to execute. The results: in the first 1000 strings of 85 digits of π , there are 535 "hits" and "465" misses.

$$\frac{535}{1000} = 0.535$$

Wowser! Interpret this, of course, with a probabilistic grain of salt. Nevertheless, there is the sense that the Law of Large Numbers is kicking in. It was a satisfying number to see appear after much toiling. I did not look even deeper (almost on aesthetic grounds) because of the computing time required by my inefficient program. I do intend to improve the program at some time in the future.

2. Barry Bonds

In 2004, Barry Bonds set the single season record for on base percentage with a staggering 0.609. For the baseball-uninitiated,

$$\text{On base percentage} = \frac{\text{hits} + \text{walks} + \text{hit by pitches}}{\text{at bats} + \text{walks} + \text{hit by pitches} + \text{sac flies}}$$

In simple terms, it (almost perfectly) measures the ratio: $\frac{\# \text{ of times reaching base safely}}{\# \text{ of times coming to bat}}$

Barry Bonds' 0.609 says that in 2004 he reached base safely more than 60% of the time. This is an amazing figure for a sport that considers a 35% rate to be strong.

I wondered: what was Bonds' longest consecutive run of reaching base safely in 2004? I didn't have access to this kind of information, but using his basic season stats, I figured I could at least make a prediction.

Here are the relevant stats for Bonds in 2004:

At bats: 373
Hits: 135
Walks: 232
Sac Flies: 3
Hit By Pitch: 9

Using the formula above, we verify: $\frac{135 + 232 + 9}{373 + 232 + 3 + 9} = \frac{376}{617} \approx 0.609$

If we assume that Bonds had a 60.9% chance of reaching base every time he stepped to the plate, then we can apply the formulas in this handout (in particular, those in #3).

For example, given these assumptions, the likelihood that Bonds reached base safely 8 times in a row at some point in the 2004 season would be

$$\text{ProbRun}(8, 0.609, 617)$$

Using the run of heads as a template, we can compute this quite easily with Excel. We find that

$$\text{ProbRun}(8, 0.609, 617) \approx 0.9921$$

Wow, what a high probability for a run of at least 8!

Let's move the number up to a run of at least 12. We compute this and determine that

$$ProbRun(12, 0.609, 617) \approx 0.4654$$

A big drop.

Okay, so how do we make a reasonable prediction for the size of the longest run in Barry Bonds' 2004 season? Expected value. We need to compute

$$ExpMaxRun(0.609, 617) = \sum_{i=1}^{617} ProbRun(i, 0.609, 617)$$

That's a lot of work even with Excel. This would require doing 617 separate spreadsheets and then summing the relevant values drawn from each. This can be done with a class of willing students by assigning 30 or so cases to each student for HW and then summing all the result after they are electronically handed in. I took the easy way out and used *Mathematica* to build a function to do the job. My result:

$$ExpMaxRun(0.609, 617) \approx 11.7288$$

Now I had a number to work with. While this number really told me about what should happen on average if Barry Bonds had something like 100,000 seasons identical to his 2004 season, I was very curious to see how the theoretical matched up with the actual.

I emailed a baseball stat site called Baseball Prospectus, explaining my "need" for this information. After several weeks I received a reply that they did not have that data readily available, but they did have a complete accounting of Barry Bonds' season, with a breakdown of each game. Super! I had no problem scrolling through the Barry Bonds' 2004 season blow-by-blow and finding that his five longest runs of reaching base safely had lengths

11, 10, 8, 8, and 7.

Perhaps it is dumb luck, but these values look very reasonable given the number-crunching I did. I was pretty excited to see the accuracy of the theoretical prediction.