Detecting Fraud and Streaks in Sports

Bruce Bukiet

Department of Mathematical Sciences
Center for Applied Mathematics and Statistics
Department of Biomedical Engineering
New Jersey Institute of Technology
Newark, NJ 07102
bukiet@m.njit.edu

for

NCTM
Eastern Regional Conference
Somerset, NJ
25-27 October 2001

Benford's Law and Detecting Tax Fraud

We can catch a large percentage of certain types of fraudulent behavior using simple mathematical ideas. If we look at the first digit of numbers that arise from data that is expected to grow exponentially over time, the 9 digits 1-9 should not appear with uniform distribution 1/9. Examples include population data, stock prices, stock market volumes.

If some process follows $\frac{dP}{dt} = kP$ then $P = P_0 e^{rt}$. Let $P_0 = 1$ (The result is not affected by this). Then, $P(0) = 1$.
$P = 2$ when $e^{rt} = 2$ or $t = \frac{\ln 2}{r}$
$P = 3$ when $e^{rt} = 3$ or $t = \frac{\ln 3}{r}$
$P = 4$ when $e^{rt} = 4$ or $t = \frac{\ln 4}{r}$
$P = 5$ when $e^{rt} = 5$ or $t = \frac{\ln 5}{r}$
$P = 6$ when $e^{rt} = 6$ or $t = \frac{\ln 6}{r}$
$P = 7$ when $e^{rt} = 7$ or $t = \frac{\ln 7}{r}$
$P = 8$ when $e^{rt} = 8$ or $t = \frac{\ln 8}{r}$
$P = 9$ when $e^{rt} = 9$ or $t = \frac{\ln 9}{r}$
$P = 10$ when $e^{rt} = 10$ or $t = \frac{\ln 10}{r}$

So the time it takes for the leading digit to get back to 1 is $\frac{\ln 10}{r}$.
The fraction of the time with first digit 1 is $\frac{\ln 2}{r}/\frac{\ln 10}{r} = \log 2 \approx 0.301$
The fraction of the time with first digit 2 is $\left(\frac{\ln 3}{r} - \frac{\ln 2}{r}\right)/\frac{\ln 10}{r} = \log 3/2 \approx 0.176$
The fraction of the time with first digit 3 is $\log 4/3 \approx 0.125$
The fraction of the time with first digit 4 is $\log 5/4 \approx 0.097$
The fraction of the time with first digit 5 is $\log 6/5 \approx 0.079$
The fraction of the time with first digit 6 is $\log 7/6 \approx 0.067$
The fraction of the time with first digit 7 is $\log 8/7 \approx 0.058$
The fraction of the time with first digit 8 is $\log 9/8 \approx 0.051$
The fraction of the time with first digit 9 is $\log 10/9 \approx 0.046$

If the numbers on a tax return do not have similar proportions (in their first digits) but, rather, are more uniform, the return might be suspect.

Modeling Runs of Heads or Tails

Let's start off with the specific example of finding the probability of getting at least 6 consecutive heads or 6 consecutive tails in a set of N flips of a fair coin. Clearly, if N is less than 6, we can have no run that long. If N is 6, there are 2 such runs (all 6 heads or all 6 are tails). If N is 7, there are 6 such runs (all 7 heads, all 7 tails, 6 heads followed by a tails, 6 tails followed by a head, a tails followed by 6 heads, or a heads followed by 6 tails). The complications start growing greatly if one desires to compute the number of combinations or flips that lead to a run of at least 6 in a given number of flips.

A brute force way to approach the problem is to allow all sequences of N flips be simulated on a computer and have the computer count how many of these sequences have a run of at least 6 consecutive heads or tails. Since the number of sequences of N flips is 2^N, the computational work grows quickly such that dealing with more than about 30 flips takes a long time. Dealing with 200 flips is just not possible in a reasonable amount of time. (Each time we add 10 flips to our list, the work increases by a factor of more than 1000).

Let us denote the number of sequences of N flips with at least 6 consecutive heads or 6 consecutive tails as C_N, the number of sequences that do not have have a run of 6 straight is $2^N - C_N$.

In order to analyze the problem, we note that there are only two ways to produce runs of at least length 6 in a set of N flips.

1. We already have a set of at least 6 consecutive heads or tails in a set of $N-1$ flips (in which case adding a heads or a tails results in two new sequences with at least a run of 6), or

2. The Nth flip is the cause of the first occurence of 6 consecutive heads or tails in the list. If it is a heads, the set of flips from the $N-6$th flip is THHHHHH. If it is a tails, the set of flips from the $N-6$th flip is HTTTTTT. In either case, there is exactly one way from each sequence of $N-6$ flips that does not yet have a run of 6 to get to a sequence of N flips whose Nth flip is the first occurence of a run of 6 consecutive heads or tails.

Thus, we have

$$C_N \;=\; 2\,C_{N-1} \;+\; 2^{N-6} \;-\; C_{N-6} \tag{1}$$

This difference equation has the initial conditions: $C_i \;=\; 0$ for $i \;=\; 0,1,...,5$, $C_6 \;=\; 2$.

If we divide C_N by 2^N, we have the fraction of the time a set of N coin flips should yield at least 6 consecutive heads or tails. However, if we run this on a computer, C_N will give overflow errors for N in the 40's or 50's, as C_N grows very quickly. It is, therefore, better to consider the fraction, call it F_N of time we have a run of at least 6 straight, i.e., $F_N \;=\; \frac{C_N}{2^N}$

$$\frac{C_N}{2^N} \;=\; \frac{2\,C_{N-1}}{2^N} \;+\; \frac{2^{N-6}}{2^N} \;-\; \frac{C_{N-6}}{2^N} \tag{2}$$

Thus,

$$F_N \;=\; \frac{2\,C_{N-1}}{2\cdot 2^{N-1}} \;+\; \frac{1}{2^6} \;-\; \frac{C_{N-6}}{2^6 2^{N-6}} \tag{3}$$

or

$$F_N \;=\; F_{N-1} \;+\; \frac{1}{2^6}\,(1 \;-\; F_{N-6}) \tag{4}$$

where $F_i \;=\; 0$ for $i \;=\; 0,1,...,5$, $F_6 \;=\; \frac{2}{2^6}$.

In general, the probability that the longest run has length at least k is

$$F_N \;=\; F_{N-1} \;+\; \frac{1}{2^k}\,(1 \;-\; F_{N-k}) \tag{5}$$

where $F_i \;=\; 0$ for $i \;=\; 0,1,...,k-1$, $F_k \;=\; \frac{2}{2^k}$. The probability that the longest run has exactly length k is found by subtracting: $F_N(k) \;-\; F_N(k+1)$, where $F_N(k)$ denotes the fraction of sets of N flips with a sequence of at least k heads or at least k tails.

Of course, for runs of length at least 2, the first flip can be anything, but the only way not to have a run of at least length 2 is for the flips to alternate. Thus, all the remaining flips are forced and have probability 1/2 so $F_N \;=\; 1-\left(\frac{1}{2}\right)^{N-1}$. The results for various values of k and N are given in Tables 1 and 2.

We see from these tables that in 200 flips of a fair coin, we can expect that a run of at least length 6 will occur with probability of more than 96%. However, for 200 flips, there is only a one in six chance that the longest run will be exactly of length six and the most probable longest length is seven.

Table 1: Probability of Runs of Length k or longer in N coin flips

Flips	**Probability Longest Run Is At Least**				**Probability Longest Run Is Exactly**			
	5	6	7	8	5	6	7	8
25	0.54963	0.29967	0.15078	0.07323	0.24997	0.148896	0.07755	0.03824
50	0.82093	0.54411	0.30892	0.16197	0.27682	0.23519	0.14695	0.08023
100	0.97169	0.80682	0.54234	0.31477	0.16487	0.26448	0.22757	0.14621
200	0.99929	0.96531	0.79929	0.54187	0.03398	0.16603	0.25742	0.22352
300	0.99998	0.99377	0.91197	0.69370	0.00621	0.08180	0.21827	0.25255
400	1.00000	0.99888	0.96139	0.79522	0.00112	0.03749	0.16618	0.25339
500	1.00000	0.99980	0.98307	0.86309	0.00020	0.01673	0.11998	0.23872

Table 2: Flips Needed to Achieve Given Probability of a Run of Length k or more

k	**Probability**					
	0.25	0.50	0.75	0.90	0.95	0.99
4	6	11	20	31	39	58
5	12	23	41	66	85	129
6	22	45	85	139	179	273
7	41	90	174	285	369	564
8	78	179	351	579	751	1150
9	152	357	705	1167	1516	2326
10	300	712	1415	2344	3048	4680

If the probabilities are not all one-half (e.g. if the coin is not fair or if we are dealing with repeats on rolls of a die) the analysis is much more complicated.

Probability of Winning a Seven Game Series

In the World Series, in Major League Baseball, the team that wins 4 games first, wins the championship. Suppose we have a model that enables us to compute the probability of each team winning each game of the series. The model might take as input the expected lineups of the teams along with player performance data, information on the starting pitchers and the team bullpens, the league or team home field advantage or other information. The output might include run distribution for each team, enabling computation of the probability of each team winning the game. It turns out, not surprisingly, that the starting pitcher has a large impact on the probability of a team winning the game so in a 7 game series, the probability of a team winning changes from game to game. In this section, we assume that we know the probability of each team winning each game and demonstrate an efficient way to compute the probability of each winning the series in 4, 5, 6 or 7 games.

Let P_i $(i = 1, 2, ..., 7)$ be the probability of team 1 winning the ith game and $Q_i = 1 - P_i$ be the probability of team 2 winning the game. We can compute the desired probabilities by brute force.

<u>Prob(team 1 wins in 4 games)</u> $= P_1P_2P_3P_4$ (3 multiplications)

<u>Prob(team 1 wins in 5 games)</u>: Team 1 wins 3 of first four and wins 5th game: $\binom{4}{3}\binom{1}{1} = 4$ ways to multiply $P_aP_bP_cQ_dP_5$ (4 multiplies times 4 terms plus 3 additions)

<u>Prob(team 1 wins in 6 games)</u>: Team 1 wins 3 of first five and wins 6th game: $\binom{5}{3}\binom{1}{1} = 10$ ways to multiply $P_aP_bP_cQ_dQ_eP_6$ (5 multiplies times 10 terms plus 9 additions)

<u>Prob(team 1 wins in 7 games)</u>: Team 1 wins 3 of first six and wins 7th game: $\binom{6}{3}\binom{1}{1} = 20$ ways to multiply $P_aP_bP_cQ_dQ_eQ_fP_7$ (6 multiplies times 20 terms plus 19 additions)

The same amount of work is necessary to handle the calculations for team 2's probability of winning. So the total amount of work is $2(3 + 16 + 50 + 120) = 378$ multiplications and $2(3 + 9 + 19) = 82$ additions. The total probability of team 1 winning the series is the sum of the probabilities team 1 wins (wins in 4 games plus wins in 5 games, etc.) for 3 more additions and probabillity of team 2 winning is 1 minus this probability.

It is more efficient to deal with the situation after 1 game at a time (See figure). After each number of games played we have games plus one numbers (at most) we keep around and use these to proceed to the next number of games.

For example, after 1 game, the probability that the series is 1-0 or 0-1 is P_1 and Q_1 respectively. We'll denote these P_{10} and P_{01}.

To get to the situation after game 2, we have
$P_{20} = P_2 P_{10}$;
$P_{11} = P_2 P_{01} + Q_2 P_{10}$ and
$P_{02} = Q_2 P_{01}$. So after 2 games we keep only these 3 numbers around. (4 multiplies and 1 addition)

To compute to the situation after game 3, we have
$P_{30} = P_3 P_{20}$;
$P_{21} = P_3 P_{11} + Q_3 P_{20}$;
$P_{12} = P_3 P_{02} + Q_3 P_{11}$; and
$P_{03} = Q_3 P_{02}$. And after 3 games we keep only these 4 numbers around. (6 multiplies and 2 additions)

Continuing in this manner we find (after 4 games):
$P_{40} = P_4 P_{30}$ Team 1 wins series in 4 games.;
$P_{31} = P_4 P_{21} + Q_4 P_{30}$;
$P_{22} = P_4 P_{12} + Q_4 P_{21}$;
$P_{13} = P_4 P_{03} + Q_4 P_{12}$; and
$P_{04} = Q_4 P_{03}$. Team 2 wins the series in 4 games. And after 4 games we keep only the 3 middle numbers around. The series has ended in the other cases. (8 multiplies and 3 additions)

After 5 games:
$P_{41} = P_5 P_{31}$; Team 1 wins series in 5 games.;
$P_{32} = P_5 P_{22} + Q_5 P_{31}$;
$P_{23} = P_5 P_{13} + Q_5 P_{22}$;
$P_{14} = Q_5 P_{13}$; Team 2 wins series in 5 games.;
We keep only the middle 2 numbers around. (6 multiplies and 2 additions)

After 6 games:
$P_{42} = P_6 P_{32}$; Team 1 wins series in 6 games.;
$P_{33} = P_6 P_{23} + Q_6 P_{32}$;
$P_{24} = Q_6 P_{23}$; Team 2 wins series in 6 games.;
(4 multiplies and 1 addition)

Table 3: Probability of Winning the World Series: 2001

Game	Home Pitcher	Visiting Pitcher	P(Yankees win)
1	Schilling	Mussina	0.531
2	Johnson	Clemens	0.380
3	Pettitte	Batista	0.565
4	Hernandez	Lopez	0.546
5	Mussina	Schilling	0.609
6	Johnson	Mussina	0.531
7	Batista	Pettitte	0.486
Series Totals in percent			
Prob Yanks in 4	6.2	Prob Dbacks in 4	5.7
Prob Yanks in 5	15.6	Prob Dbacks in 4	9.5
Prob Yanks in 6	12.6	Prob Dbacks in 4	18.4
Prob Yanks in 7	15.5	Prob Dbacks in 4	16.4
Prob Yanks Total	49.9	Prob Dbacks Total	50.1
Probabilities if Yankees win game 1			
Prob Yanks in 4	11.7	Prob Dbacks in 4	0.0
Prob Yanks in 5	23.1	Prob Dbacks in 4	4.8
Prob Yanks in 6	14.5	Prob Dbacks in 4	13.9
Prob Yanks in 7	15.6	Prob Dbacks in 4	16.5
Prob Yanks Total	64.9	Prob Dbacks Total	35.1

Finally, after 7 games, we have:
$P_{43} = P_7 P_{33}$; Team 1 wins series in 7 games;
$P_{34} = Q_7 P_{33}$; Team 2 wins series in 7 games.;
(2 multiplies)

So the total number of operations for the entire computation is 30 multiplications and 9 additions or about one-tenth the work of the brute force method, so this streamlined method is more likely to be performed without careless errors.

The total probability of team 1 winning the series is the sum of the probabilities team 1 wins (wins in 4 games plus wins in 5 games, etc.) for 3 more additions and probabillity of team 2 winning is 1 minus this probability.

See an example for the 2001 World Series in Table 3.

The Hot Hand

What is the appropriate way to measure whether a player has the hot hand? One simple measure was used by Tversky and Gilovich, who studied how players performed in a sequence of shots in a game after 1, 2, or 3 misses and after 1, 2, or 3 shots hit.

If the "Hot Hand" is real, we would find that players performance improves after shots made. Here is a summary of data from their paper. The performance numbers should improve as we read across the page left to right. They don't. The results were similar for individual players.

$P(H/3M)$	$P(H/2M)$	$P(H/1M)$	$P(H)$	$P(H/1H)$	$P(H/2H)$	$P(H/3H)$	*correlation*
0.56	0.53	0.54	0.52	0.51	0.50	0.46	-0.039

Larkey et al. gave a more complicated measure of the hot hand that involved how close in time and "game context" of the shots taken. Their measure DID yield support for a "hot hand" in some players.

References

Browne, Malcolm, *Following Benford's Law, or Looking Out for No. 1* in the New York Times, Tuesday, August 4, 1998.

Schilling, Mark F., *The Longest Run of Heads* College Mathematics Journal, Vol. 21, No. 3, pp. 196-207, May 1990.

Bukiet, B., Harold, E., and Palacios J., *A Markov Chain Approach to Baseball* Operations Research, Vol. 45, No. 1, pp. 14-23, Jan/Feb 1997.

For more on the baseball method, see www.egrandslam.com or http://m.njit.edu/∼bukiet

Tversky, Amos and Gilovich, Thomas, *The Cold Facts about the "Hot Hand" in Basketball* Chance, Vol. 2, No. 1, pp. 16-21, 1989

Larkey, Patrick, D., Smith, Richard A. and Kadane, Joseph B., *It's Okay to Believe in the "Hot Hand"* Chance, Vol. 2, No. 4, pp. 22-30, 1989