The Mathematics of Winning Streaks

Erik Leffler
lefflererik@gmail.com

under the direction of
Prof. Henrik Eriksson
Department of Computer Science and Communications
Royal Institute of Technology

Research Academy for Young Scientists
July 8, 2015

Abstract

The purpose of this paper is to develop mathematical tools that can investigate sequences of random numbers and how they form records. We are specifically interested in the kind of records that can be categorised as "winning streaks", records of consecutive wins. We examine the frequency of records in a series of real numbers, $x_1, x_2, x_3 \cdots$ where any x_i may be 1 with the probability of p and 0 with the probability of $1-p = q$. A record is, in this paper, defined as a streak of 1s that is longer than any of the previous streaks. Formulas are deduced for the probability that the first record ends at any given index and that any given nth record is of length l. Expected values are then calculated. These formulas can be used to analyse if records in sequences are the results of randomness or indicate underlying trends.

Contents

1 Introduction

Almost every day we hear about new records in the news. It may be that a stock index has reached an all time high or that a record low temperature has been measured in Sweden. It is in many cases of utmost importance to know whether or not these records indicate a trend or if they are just a result randomness. By investigating a sequence of random numbers and how they form records we will develop mathematical tools to examine these records.

In this paper, a specific type of record, namely "winning streaks" will be investigated. A football team, a politician, an athlete or competitor in a win-or-lose-activity, all count consecutive wins. Sometimes such records indicate a better team, athlete or politician. Sometimes they just indicate randomness.

There are many different observations to be considered regarding these winning streaks. What is the probability that the nth winning streak is of length l? What is the expected length of the nth winning streak? What is the probability that the first winning streak ends after e tries? Where is the first record expected to end? These problems will be analysed with mathematics and checked by the use of Python simulations.

The area of analysing record sequences by the use of mathematics is relatively unexplored. Therefore, no earlier papers are being referred to. A similar paper by Jonna Karlberg [1] (publication pending) analyses temperature records but I know of no earlier papers on winning streak records. Any textbook on probability theory may be used as a reference for the elementary formulas used here.

1.1 Notation and Definitions

Consider a sequence of numbers, $x_0, x_1, x_2, x_3 \cdots$ where any x_i may be 1 with a probability of p or 0 with a probability of $q = 1 - p$. An x_i is considered a win if it is equal to 1. A winning streak is a sequence of consecutive wins followed by a loss. A record is a winning streak that is longer than any previous winning streak. The entire series can be divided into smaller sequences that we call puzzle pieces. A puzzle piece consists of a potential losing streak, followed by a winning streak and ends with a loss. One example of a puzzle piece would be [010], which would appear with a probability of $p \cdot q^2$. Another puzzle piece could be [00001110], which would appear with a probability of $q^4 \cdot p^3 \cdot q$. However, [00100110] would not be a puzzle piece as it has two streaks of consecutive wins and could therefore be split up into the pieces, [0010] and [0110]. As shown above, one piece must contain one and only one streak of consecutive wins. A combination of puzzle pieces will be referred to as a puzzle. Table 1 shows the probability of each different puzzle piece. In the first column we see the probabilities for pieces [10], [010], [0010] etc.

Table 1: Probability table

$q^1 \cdot p^1$	$q^1 \cdot p^2$	$q^1 \cdot p^3$	$q^1 \cdot p^4$	$q^1 \cdot p^5 \cdots$
$q^2 \cdot p^1$	$q^2 \cdot p^2$	$q^2 \cdot p^3$	$q^2 \cdot p^4$	$q^2 \cdot p^5 \cdots$
$q^3 \cdot p^1$	$q^3 \cdot p^2$	$q^3 \cdot p^3$	$q^3 \cdot p^4$	$q^3 \cdot p^5 \cdots$
$q^4 \cdot p^1$	$q^4 \cdot p^2$	$q^4 \cdot p^3$	$q^4 \cdot p^4$	$q^4 \cdot p^5 \cdots$
$q^5 \cdot p^1$	$q^5 \cdot p^2$	$q^5 \cdot p^3$	$q^5 \cdot p^4$	$q^5 \cdot p^5 \cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

From here on puzzle pieces will be referred to in this way, for instance, $q^3 \cdot p^3$ is how the puzzle piece [001110] will be referred to. A sequence must start with exactly one of the puzzle pieces, so the probabilities of every puzzle piece should add up to one. The lth row is a geometric series with the sum,

$$\sum_{k=1}^{\infty} p^k \cdot q^l = \frac{q^l \cdot p}{1-p} = p \cdot q^{l-1}.$$

By adding the sums of every row we get,

$$\sum_{l=1}^{\infty} p \cdot q^{l-1} = \frac{p}{1-q} = \frac{p}{p} = 1.$$

As expected, the sum of all puzzle pieces is indeed equal to one.

The end of a record is defined to be the index of the 0 that breaks off the record. The length of a record is defined to be the amount of consecutive 1s. We let E_n be the stochastic variable that denotes where the nth record ends. L_n is the stochastic variable that denotes the length of the nth record. $E(i)$ denotes the expected value of i. $P(\sigma)$ denotes the probability that a statement σ is true.

2 Calculations and Results

2.1 Calculating the expected ending of the first winning streak

Here we want to calculate the probability of the first record ending at location e. We denote this as $P(E_1 = e)$. Seeing as every puzzle piece must contain one and only one record, and that a puzzle piece and a record end at the same place (the 0 that breaks off the record), $P(E_1 = e)$ can be interpreted as the probability that the total length (the amount of 0s and 1s) of the first puzzle piece is equal to e. Therefore we want to sum the probability for every puzzle piece where the amount of 0s and 1s add upp to e i.e, $p^{e-k} \cdot q^k$ for every k so that $1 \leqslant k < e$. We get the following sum,

$$P(E_1 = e) = \sum_{k=1}^{e-1} p^{e-k} \cdot q^k = \sum_{k=1}^{e-1} p^e \cdot \left(\frac{q}{p}\right)^k = p^e \cdot \frac{q}{p} \cdot \frac{1 - \left(\frac{q}{p}\right)^{e-1}}{1 - \frac{q}{p}} = \frac{q}{p} \cdot \frac{p^e - p \cdot q^{e-1}}{\frac{p-q}{p}} = \frac{q \cdot p^e - p \cdot q^e}{p - q}. \tag{1}$$

Since $P(E_1 = e) = \frac{q \cdot p^e - p \cdot q^e}{p-q}$ we can calculate the expected value of E_1, $E(E_1)$, as

$$E(E_1) = \sum_{e=1}^{\infty} e \cdot \frac{q \cdot p^e - p \cdot q^e}{p-q} = \frac{q \cdot p}{p-q} \sum_{e=1}^{\infty} e \cdot p^{e-1} - e \cdot q^{e-1} = \frac{q \cdot p}{p-q} \left(\sum_{e=1}^{\infty} e \cdot p^{e-1} - \sum_{e=1}^{\infty} e \cdot q^{e-1} \right). \tag{2}$$

The two sums in equation (2), $\sum_{e=1}^{\infty} e \cdot p^{e-1}$ and $\sum_{e=1}^{\infty} e \cdot q^{e-1}$ can be calculated by using the derivative of the geometric series,

$$1 + 2 \cdot x + 3 \cdot x^2 + 4 \cdot x^3 + \cdots = \frac{1}{(1-x)^2} \tag{3}$$

and thus

$$E(E_1) = \frac{q \cdot p}{p-q} \cdot \left(\frac{1}{(1-p)^2} - \frac{1}{(1-q)^2} \right) = \frac{1}{p-q} \cdot \left(\frac{p}{q} - \frac{q}{p} \right) = \frac{1}{p-q} \cdot \left(\frac{p^2 - q^2}{qp} \right) = \frac{p+q}{qp} = \frac{1}{q} + \frac{1}{p} \tag{4}$$

.

The value of this is verified by running a Python script which can be found in the appendix.

Something very interesting about this result is that $P(E_1 = e)$ is symmetrical with respect to p and q. If the value p were to be swapped with q then $P(E_1 = e)$ would remain the same. A practical interpretation of this would be that a very good basketball player that scores 90% of the time sets his first record after the same amount of throws as someone who scores 10% of the time.

2.2 Calculating the expected length of the first winning streak

We want to calculate the probability that the first record is of length l. We denote this as $P(L_1 = l)$ where L_1 is equal to the length of the first record. We calculate $P(L_1 = l)$ by summing the probabilities of getting any puzzle piece with a winning streak of length

l. Any puzzle piece with length l appears with the probability $q^i \cdot p^l$ for any $i \geqslant 1$ The summation of all of these probabilities can be expressed as:

$$P(L_1 = l) = \sum_{i=1}^{\infty} p^l \cdot q^i = q \cdot p^l \cdot \frac{1}{1-q} = q \cdot p^{l-1}. \tag{5}$$

This can be visualised as the sum of the lth column in table 1. This can also be interpreted as the probability that any puzzle piece contains a wining streak that is of length l.
Now we want to calculate the expected value of the length of the first record. We denote this as $E(L_1)$. This will be equal to the probability that a certain length l appears multiplied with the length, l, itself. Thus,

$$E(L_1) = \sum_{l=1}^{\infty} l \cdot q \cdot p^{l-1} = \frac{q}{(1-p)^2} = \frac{1}{q}. \tag{6}$$

The value of this is verified by running a Python script which can be found in the appendix.

2.3 Calculating the expected length of the nth record

We want to calculate the probability that the nth record is of length l. We call this $P(L_n = l)$ where L_n is the length of the nth record. We do this by adding the probability of every single puzzle that meets the condition that the nth record is of length l.
Initially, we fix the mth and $m+1$th record at two given lengths, L_m and L_{m+1}. We do this in a given sequence so that the condition that $L_n = l$ is true. We examine all possible puzzles that lie in between the mth and $m+1$th record. It is not allowed for any puzzle piece between the mth and $m+1$th record to have a record length greater than L_m. If the length would be greater than L_m then that puzzle piece would contain the $m+1$th record and could therefore not lie between the mth and $m+1$th record. Thus we want to calculate the probability, $P(L \leqslant L_m)$ for any given allowed puzzle piece. The probability that any puzzle piece has the record length of L is equal to $q \cdot p^{L-1}$ (see subsection: 2.2). Due to the fact that any puzzle piece between the mth and $m+1$th record is allowed to

have a length L where $1 \leqslant L \leqslant L_m$ we get that

$$P(L \leqslant L_m) = \sum_{L=1}^{m} q \cdot p^{L-1} = q \cdot \frac{1 - p^{L_m}}{1 - p} = 1 - p^{L_m}.$$

Now that we know the probability for any given allowed puzzle piece between the mth and $m + 1$th record, we want to calculate the probability for any given amount of these pieces. We sum the probability for obtaining zero pieces, one piece, two pieces$\cdots$ and get,

$$\sum_{i=0}^{\infty} (1 - p^{L_m})^i = \frac{1}{1 - (1 - p^{L_m})} = \frac{1}{p^m}.$$

Calling this a probability is, of course, an error. In our case this is going to be a factor in our final probability. $\frac{1}{p^m}$ will always be larger than or equal to one. If this calculation was to be done correctly an ending to every non-record puzzle would have to be specified. If we were to give an ending to every puzzle we would do this by multiplying every non-record puzzle with the probability that a record is drawn. The probability that a puzzle piece contains a record larger that L_m is equal to,

$$P(L > L_m) \cdot \sum_{L=L_m+1}^{\infty} q \cdot p^{L-1} = q \cdot p^{L_m} \cdot \frac{1}{1 - p} = p^{L_m}.$$

We see that if we multiply $P(L_k > L_m)$ with every non-record puzzle piece we get,

$$\sum_{i=0}^{\infty} p^{L_m} \cdot (1 - p^{L_m})^i = p^{L_m} \cdot \frac{1}{1 - (1 - p^{L_m})} = 1.$$

And the sample space is equal to one. However, in our case we are going to multiply the non-ending probability factor with the probability of getting a record of length L_m, which is going to be the record previous to the non-record streak. We get, $\frac{1}{p^m} \cdot q \cdot p^{m-1} = \frac{q}{p}$ We call this P_{f,L_m}, the probability factor of getting a record of length L_m followed by every possible non-record puzzle. If we then multiply together $n - 1$ of these then every P_{f,L_m} is going to be followed by a $P_{f,L_{m+1}}$ for all values of m except $m = n - 1$. Due to the fact

that the start of a probability factor is going to specify an end to the previous probability factor we have specified an ending to every probability factor except the last one, $P_{f,L_{n-1}}$. Later we are going to end the final non-record puzzle with the nth record. We want to get $n-1$ probability factors, the probability of this is, $\left(\frac{q}{p}\right)^{n-1}$. The fact that we want $n-1$ probability factors and have $l-1$ to choose from gives us that $\binom{l-1}{n-1} \cdot \left(\frac{q}{p}\right)^{n-1}$ is the probability of getting any $n-1$ probability factors. If we then multiply this with the probability, $P(L_n = 1)$, which is finally going to give us a last ending and a true probability, then we get,

$$q \cdot p^{l-1} \cdot \binom{l-1}{n-1} \cdot \left(\frac{q}{p}\right)^{n-1} = \binom{l-1}{n-1} \cdot q^n \cdot p^{l-n}$$

which is going to be equal to $P(L_n = l)$. Hence, we get,

$$P(L_n = l) = \binom{l-1}{n-1} \cdot q^n \cdot p^{l-n} \tag{7}$$

We can use this to now calculate the expected value of L_n, $E(L_n)$. $E(L_n)$ is going to be equal to $P(L_n = l) \cdot l$ for every possible l. We get the following sum,

$$\sum_{l=n}^{\infty} l \cdot \binom{l-1}{n-1} \cdot q^n \cdot p^{l-n} =$$
$$\sum_{l=n}^{\infty} l \cdot \frac{(l-1)!}{(n-1)! \cdot (l-n)!} \cdot p^l \cdot \left(\frac{q}{p}\right)^n =$$
$$\sum_{l=n}^{\infty} \frac{(l)!}{(n-1)! \cdot (l-n)!} \cdot p^l \cdot \left(\frac{q}{p}\right)^n =$$
$$\frac{1}{(n-1)!} \cdot \left(\frac{q}{p}\right)^n \cdot \sum_{l=n}^{\infty} p^l \cdot \frac{l!}{(l-n)!}$$

To calculate $\sum_{l=n}^{\infty} p^l \cdot \frac{l!}{(l-n)!}$ the nth derivative of the geometric series is needed. Let

$$f(x) = x^m \rightarrow f^{(n)}(x) = \frac{m!}{(m-n)!} \cdot x^{m-n}$$

Let

$$S(x) = x^1 + x^2 + x^3 + \cdots = \frac{1}{1-x}$$

$$\rightarrow$$

$$S^{(n)}(x) = \frac{(n+0)!}{0!} \cdot x^0 + \frac{(n+1)!}{1!} \cdot x^1 + \frac{(n+2)!}{2!} \cdot x^2 + \frac{(n+3)!}{3!} \cdot x^3 + \cdots = \frac{n!}{(1-x)^{(n+1)}} \quad (8)$$

The important thing to deduce from this equation is that,

$$\frac{n!}{(1-x)^{(n+1)}} = \sum_{l=n}^{\infty} \frac{l!}{(l-n)!} \cdot x^{(l-n)}. \quad (9)$$

If we use this relationship to calculate $E(L_n)$ we get,

$$E(L_n) = \frac{1}{(n-1)!} \cdot \left(\frac{q}{p}\right)^n \cdot \sum_{l=n}^{\infty} p^l \cdot \frac{l!}{(l-n)!} = \frac{1}{(n-1)!} \cdot \left(\frac{q}{p}\right)^n \cdot p^n \cdot \frac{n!}{(1-p)^{(n+1)}} = \frac{n}{q} \quad (10)$$

This value is verified using Python code found in the appendix .

3 Applications

The results deduced in this paper can be used to analyse different winning streak sequences. If the formulas give results that are somewhat alike the measurements that are being analysed then it is fair to assume that p and q are constant in the analysed sequence. If they differ then that might be due to a change of p and q which in turn might indicate an underlying trend, i.e a basketball player improving his talents or a market crisis.

4 Acknowledgements

First of all I would like to thank my mentor and supervisor Henrik Eriksson. Being able to discuss the problems of this report with him has been invaluable. Jonna Karlberg who I have been working beside has also been valuable in discussions concerning the project. Finally I would like to thank Research Academy for Young Scientists for allowing me to participate in their program.

References

[1] Karlberg J. Analysis of Temperature Records in Random Series. July 8, 2015

A Appendix

A.1 Simulation code for $E(E_1)$

```
from random import random



sum = 0
p=0.1 #Adjust for desired value of p
n=1000 # Adjust for desired amount of runs
for k in range(n):
    counter = 0
    s=0
    while 1 > 0:
        a = random()
        if a <= p:
            sum = sum + 1
            s=1


        else:
            sum = sum + 1
            if s == 1:
                 break



    mean = sum/float(n)


print mean
```

A.2 Simulation code for $E(L_n)$

```
from random import random


def run():
    Clist = list()
    p=1-0.5 #Adjust for different desired values of p
    counter=0
    Clist.append(counter)
    s=100 #Adjust to satisfy desired value of n

    for k in range (0,10):

        while 1 > 0:
            a=random()
            if a<=p:
                counter += 1


            else:
                if counter > max(Clist):
                    Clist.append(counter)
                    counter = 0
                if len(Clist)>s:
                    break


    return Clist[100]


sum = 0
```

```
n = 500 #Adjust to satisfy amount of runs
for i in range(n):
    sum = sum + run()

sum = sum / n
print(sum)
```