Article Summary of The Suprising Predictability of Long Runs by Mark F. Schilling

Gabriel Wies

April 2021

Abstract

In the natural world we encounter many events comprising of independent Bernoulli trails such as tossing a coin or roulette. While it may seem counter-intuitive, in these events we are able to predict the length of the longest run of identical outcomes with surprising accuracy. Mark F. Schilling's paper on this topic provides equations that can predict the length along with an approximation of the accuracy of these predictions.

1 The approximate length of the longest success run

When events can be modeled as independent Bernoulli trails, the length of the longest run can be represented by L. Suppose we have a sequence of n Bernoulli trails with a probability of p and $q = 1 - p$, we estimate longest run with the function:

$$l = \log_{\frac{1}{p}}(nq) \tag{1}$$

We can derive this function by defining the random variable $L = L_n$ to the length of the longest success run. We can set the expected value of the number of failures (N_F) equal to nq where, as defined earlier, $q = 1 - p$. A fraction of p^n failures should be followed by at least n successes, therefore it should average out to nqp^l runs of l length successes. In order to find the longest possible length of l it is logical to set the number of occurrences equal to 1 (i.e. $nqp^l = 1$). Solving for l obtains Eq.1.

Using Eq.1 we can estimate the longest run of heads when flipping a coin $(p = \frac{1}{2})$. The values of l have been rounded to the nearest integer.

No. of trials (n)	Head runs (l)
100	6
1000	9
10,000	12
1,000,000	19

2 The longest run of identical outcomes

The model as defined so far can is only applicable to the longest run of Bernoulli trials. This, however has limited uses and we are primarily interested in the longest run of identical outcomes rather than the longest run of a specific outcome. For example, we could be looking for the longest run of any side of a six-sided die. In this case we are less interested in the number the die lands on, but rather how long the streak is. We call trials with several equally likely outcomes multinomial trials. In order to adapt this model for multinomial trials we can call an outcome a "succes" if it is a repeat of the previous outcome. A sequence of l "successes" is identical to $l + 1$ repeating values. Consider a fair die experiment where we set the possible outcomes to $1, 2, 3, 4, 5, 6$.

Side rolled	1	3	3	4	5	5	5	5	6	2	1
Repeat?	-	N	Y	N	N	Y	Y	Y	N	N	N

Since the sequences can be represented by Y and N, we can still apply the Eq.1 to a multinomial model. In a multinomial trial, the number of consecutive successes are identically distributed and independent.

3 Predictability of the length of the longest run

A prediction such as Eq.1 is just an estimate of the likely length of L_n in a Bernoulli trial. Since the number of consecutive successes are identically distributed and independent, we can form a probability of mass function $p(k) = p^k q$ due to the geometric distribution of the random variables. The distribution of L_n is therefore approximated to be the maximum of $[nq]$ independent geometric random variables. Schilling proves this finding through exponential distribution:

Let X be and exponential random variable with $\lambda = -\ln p$ as a parameter. The density function is given by

$$\lambda = e^{-\lambda x} = \lambda p^x$$

for $x \geq 0$. Let $Y = [X]$; then

$$P(Y = k) = \int_k^{k+1} e^{-\lambda x} dx = e^{-\lambda x}(1 - e^{-k}) = p^k q$$

In order to approximate the estimation error of l, we let $X_1, X_2, X_3,...$be independent exponential random variables each with parameter $\lambda = -\ln p$, and let $M_n = max(X_1, X_2, X_3, ..., X_{[nq]})$. Since the value of M_n tends to increase as n increases, it is better to look at the difference between M_n and our estimate using Eq.1. This difference can be denoted by

$$E_n = M_n - \log_{1/p}(nq)$$

Schilling sets F_E equal to the limiting cumulative distribution function of E_n in order to derive the limit of the probability of the estimation error denoted by $F_E(x)$

$$F_E(x) = \lim_{n \to \infty} P(E_n \le x) = e^{-p^x} \tag{2}$$

E_n therefore has a limiting distribution given by Eq.2, which is said to have an extreme value distribution known as Gumbel distribution. The mean and standard deviation are as follows

$$\mu = \gamma \ln\left(\frac{1}{p}\right)$$

$$\sigma = \frac{\pi}{\sqrt{6}\ln\left(1/p\right)} \tag{3}$$

It can be concluded;

$$P(L_n = l) \approx P(\log_{1/p}(nq) \le X < l + 1 - \log_{1/p}(nq)) \tag{4}$$

From the result of Eq.4 it can be concluded that the approximate distribution of the prediction error $L_n - \log_{1/p}(nq)$ does not depend on n. That means that it is possible to predict the longest run in any length trial as long as $nq \gg 1$. The standard deviation formula (3) shows that the spread of distributions is generally small, implying great predictability in the length of the longest run.

For example in $100,000,000$ spins of American Roulette ($p = 18/38$) the prediction interval for the longest run of black or red with an approximate 95% prediction interval will be $(22, 28)$ while the prediction interval for $500,000,000$ spins is $(24, 30)$. Note how the prediction interval length does not grow even when n is increased 5 times over.

4 Extremely predictable cases

The approximating extreme value density becomes more concentrated as the value of p becomes smaller, and therefore the length of the longest success run becomes even more predictable. In cases where p is small, we can almost always guarantee the length of the longest success run. Schilling shows this by setting N_l equal to the number of success runs of length l or longer in n Bernoulli trails. If we treat n_F as a given, we obtain

$$P(L_n = l) = P(N_l > 0, N_{l+1} = 0 | n_F) \approx e^{-n_F p^{l+1}} - e^{n_f p^l}$$

By the Law of Large Numbers, if n is large we can treat n_f as equal to its expected value of nq. Schilling solves $e^{-nqp^l} = p^{1/q}$ for n yielding

$$n = \frac{\ln(1/p)}{q^2 p^l} \tag{5}$$

which gives

$$P(L_n = l) \approx p^{p/q} - p^{1/q} \tag{6}$$

Eq. 5 solves for the values of n that maximize $P(L_n = l)$ for a given p and run length of l. Schilling also derives that the limit of Eq. 6 as p approaches 0 is equal to 1. This shows that the longest success run of small value of p is likely to have a length of l. As an example, Schilling looks at the first $5,000,000$ digits of π. This value of n is near optimal for predicting the run length of any two digit pair $\left(p = \left(\frac{1}{10}\right)^2\right)$. Plugging these values into Eq.6 it can be said with 94.5% probability that the longest run of any two-digit pair will be three. Looking at the first $4,700,000$ digits of π, the longest run for 94 out of 100 of these pairs is 3. Values of n that satisfy Eq.5 approximately place almost all of their mass of the probability distribution of the longest run on a single value.

References

[1]   Mark F. Schilling. "The Surprising Predictability of Long Runs". In: *Mathematics Magazine* 85.2 (2012), pp. 141–149.