Ad fontes.
The surprising predictability of long runs

Sophia Kholod, Sophia Zhyrovetska

December 2018

1 Problem statement

Lottery, roulette, coin tossing, digits of pseudorandom numbers - turns out that these, at first sight totally incomparable and unpredictable sequences, have one thing in common - they all are quite predictable. Turns out, that we can predict very accurately the longest run in all these sequences - and quite easily. The model described in this paper is very flexible - it can fit to any of these sequences of any length and any success probability (if it's a sequence of independent Bernoulli trials). We also provide an example of Florida Pick 2 Midday Lottery for better understanding of the model.

2 Theoretical part

The length of the longest run

If we have a sequence of independent Bernoulli trials of length n, we can define a random variable L_n as length of the longest success run. The sequence itself has a binomial distribution $\beta(n, p)$, where p - the probability of a success in Bernoulli experiment(respectively, the probability of failure is $q = 1 - p$). The expected value of failure in one trial is equal to q, whereas in n trials $= n \cdot q$. After each failure, we can expect a success run, so, roughly, there will be $n \cdot q \cdot p^l$ success runs of length l. As we want to maximize the length of it, we can say that we're looking for the only longest run. Hence, by solving an equation $nqp^l = 1$ we obtain that the length of it should be as follows:

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

If we need to consider not only success runs, but any repeating pattern in multinomial trials, we can use the following approach. We denote an outcome as a success when it repeats the previous one. Now a sequence of l successes is the same as a sequence of $l + 1$ consecutive identical values. According to the constructed model, a sequence of l successes means sequence of length $l + 1$ of identical values. Thus, the length of it will be $l = log_{1/p}(nq) + 1$.

The distribution of L_n

The formula (1) is our "best guess", but we can estimate an error to make the prediction better. To make it, we need to find the distribution of L_n. Since the success run ends when we encounter a failure, its probability distribution is geometric:

$$p(L_n = k) = p^k q$$

for $k = 0, 1, \ldots, \lfloor n \cdot q \rfloor$. The geometric distribution is discrete, thus, the limit of the maximum of L_n does not exist. But we can transfer to its continuous equivalent - exponential distribution. Let X be an exponential r.v. with parameter $\lambda = -\ln p$, hence, its density is:

$$f(x) = \lambda e^{-\lambda x} = \lambda p^x \text{ for } x > 0$$

Thus,$P(\lfloor X \rfloor = k) = \int_k^{k+1} \lambda e^{-\lambda x} dx = p^k q,$ which is precisely the geometric p.m.f.

Let E_n be the prediction error, then: $E_n = M_n - log_{1/p}(nq)$, where $M_n = max(X_1, X_2, \ldots, X_{\lfloor nq \rfloor})$. Its cumulative distribution function is

$$\begin{aligned} F_E(x) &= lim_{n\to\infty} P(M_n - log_{1/p}(nq) \leq x) = \\ &= lim_{n\to\infty} [P(X_1 \leq log_p 1/nq + x)]^{\lfloor nq \rfloor} = \text{by the c.d.f. of } X_1 \\ &= lim_{n\to\infty} [1 - p^{log_p 1/nq + x}]^{\lfloor nq \rfloor} = \\ &= lim_{n\to\infty} [1 - \frac{1}{nq} p^x]^{\lfloor nq \rfloor} = \\ &= e^{-p^x} \end{aligned} \tag{2}$$

which the c.d.f. for extreme value distribution.
We have established that the distribution of $L_n - log_{1/p}(nq)$, the length of the longest success run in n independent Bernoulli trials minus its predicted value given by (1), is well approximated by $\lfloor E_n \rfloor$, which distribution is given in (2). Therefore

$$P(L_n = l) \approx P(l - log_{1/p}(nq) \leq E < l + 1 - log_{1/p}(nq))$$

This result shows that approximate distribution of E_n does not depend on n, so we can predict the longest run in hundred trials as well as in million. Also, the spread of extreme value distribution is not that wide, so our prediction will be quite accurate. The code of the algorithm is on github [3].

In situations, when p is very close to 0, we can predict the length of the longest run with even better accuracy. The probability that the run of successes has at least length l is p^l, and if n_F is the number if failures, the desired probability is

$$P(L_n = l) = (1 - p^{l+1})^{n_F} - (1 - p^l)^{n_F} \approx e^{-n_F p^{l+1}} - e^{-n_F p^l}.$$

Maximizing the probability and using the fact that n_F approaches its expected value (by the LLN), we get

$$P(L_n = l) \approx p^{p/q} - p^{1/q}. \tag{3}$$

3 Real life application

Once a day a Florida Pick 2 Midday Lottery[2] runs a game, in which two numbers are chosen randomly every day. So, the probability of success is 0.01 - p is quite small, so we can use formula (3) for prediction. The calculations (details are in R notebook) yield the result of $l = 2$ with probability 0.945.

```
longest_run_len_prediction(0.01, 500)
```

```
[1] 1.013497
```

With probability:

```
find_probability(0.01)
```

```
[1] 0.945003
```

Analyzing the data, we got that the longest run is indeed equal to 2.

	Date <fctr>	Winning.numbers <fctr>
24	Nov 21, 2018	0 3
25	Nov 20, 2018	6 4
26	Nov 19, 2018	6 4
27	Nov 18, 2018	7 4

4 rows

	Date <fctr>	Winning.numbers <fctr>
36	Nov 09, 2018	1 2
37	Nov 08, 2018	3 6
38	Nov 07, 2018	3 6
39	Nov 06, 2018	5 0

4 rows

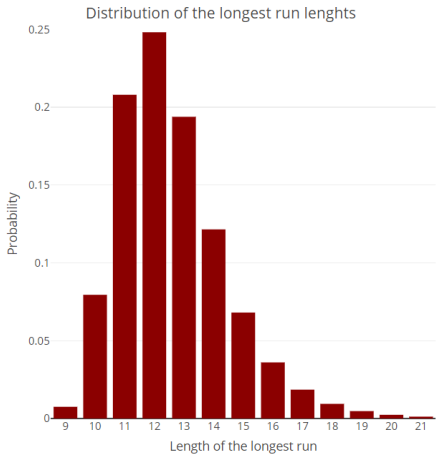
```
[1] 2
```

The main application of our prediction is that if we got no consecutive games with the same winning numbers, or if we got the length of it bigger than 2, we can predict with great probability that the lottery was falsified.

4 Conclusions

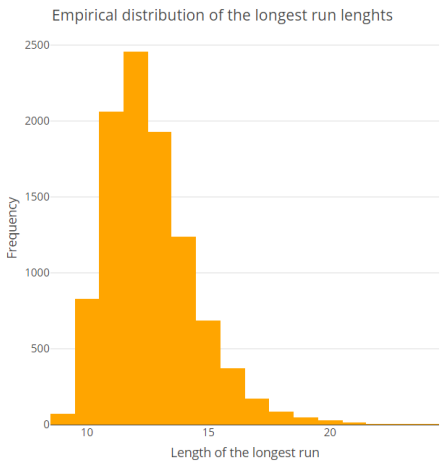
We have conducted experiments to justify the derived formulas. In this experiment, we calculated the distribution of L_n, and then compared its c.d.f. to the empirical c.d.f. - we simulated the fair coin tosses and then calculated the length of the longest run.

```
p = 1/2
n = 10000
longest_run_distribution(p,n)
```



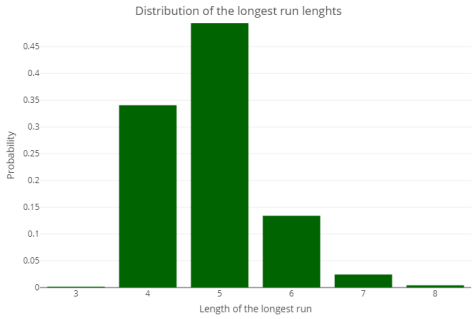
Our experiments justified that the bigger n is, the bigger is the probability that its length will be equal to our rule of thumb estimate +- 1. For this case, $log_{1/p}nq \approx 12.3$, and the 3 most frequent values are $11, 12$ and 13. Empirical c.d.f. also justifies it.

```
longest_run_empirical_distribution(n)
```

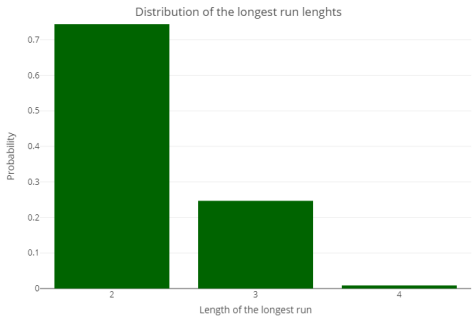


Applying the knowledge of this distribution to the real world data, we can check if the lottery was falsified or no. Furthermore, we can use the same approach not only for random games, but also for calculating the repeating patterns in pseudo random sequences - mathematical constants such as π, e, ϕ(golden ration) and so on.

We have also checked the statement that the smaller p is, the better prediction we can make (3). We have conducted an experiment for $p = 1/6$ and $n = 10000$ and got this c.d.f.



And this is the c.d.f. for $p = 1/32$ and $n = 10000$



As can be seen, the difference of the probability for the longest run is tremendous, thus, it justifies the statement.

References

[1] MARK F. SCHILLING. *The Surprising Predictability of Long Runs*. California State University, Northridge.

[2] Florida Pick 2 Midday Lottery `https://www.lotterycorner.com/fl/pick-2-midday/2018`

[3] `https://github.com/lazyTurtle21/Long-runs`