A Puzzle about Random Sequence

Yanhua Huang

Dec 2017

Let $\mathbf{X} = (X_1, X_2, ..)$ be a sequence of iid discrete random variables with $p_i = P(X_j = i)$. For a given pattern $i_1, ..., i_n$ let $T = T(i_1, ..., i_n)$ denote the number of random variables that we need to observe until the pattern appears. How to calculate $\mathbb{E}[T]$?

Let's first consider whether the pattern has an overlap which means that for some $1 \leq k \leq n$ and any $j = 1, 2, ..., k$ we have $i_j = i_{n-k+j}$.

Case 1: No overlaps

First of all, we can simply get $T = j + n$ if and only if the pattern does not occur within the first j values and $(X_{j+1}, ..., X_{j+n}) = (i_1, ..., i_n)$, i.e.,

$$P(T = j + n) = P(T > j, X_{j+1:j+n+1} = i_{1:n+1}). \tag{1}$$

However, $T > j$ is only determined by the first j values. Consequently,

$$P(T = j + n) = P(T > j) \prod_{j=1}^{n} p_{i_j}. \tag{2}$$

Summing both sides over all j yields,

$$1 = \sum_{j=0}^{\infty} P(T = j + n) = \sum_{j=0}^{\infty} P(T > j) \prod_{j=1}^{n} p_{i_j}. \tag{3}$$

Combining with $P(T > j) = \sum_{k=1}^{\infty} P(T = j + k)$ we can get $\mathbb{E}[T] = (\prod_{j=1}^{n} p_{i_j})^{-1}$.

Case 2: Has Overlaps

We can add an element at the tail to get a new pattern without overlap. Then we can get the solution by reusing the discussions in case 1.