

Explicit formulae for distributions of words and their computational complexity

Hayato Takahashi

Random Data Lab. Inc.

Sep. 7, 2023
IIBMP 2023, Tokyo Univ.

Problem: counting words in finite strings

The number of the occurrences of words in finite strings plays important role in information theory, genome analysis, statistics, AI, etc.

Example: The words 10 and 00 appear in 100010010 three times.

Explicit formulae for distributions of words are only known for several types of counting of runs (e.g. 00).

Those for nonoverlapping words (e.g. 10) are even unknown!

Known results, generating functions

The probability of the number of occurrences of words given as rational generating functions (e.g. Bassino et.al [4], Regnier et.al [17], and Robin [18]).

$f(n, k, w)$: probability that w appears k times in sample size n . Then

$$\sum_{n,k} f(n, k, w) z_1^n z_2^k = \frac{g(z_1, z_2)}{h(z_1, z_2)}.$$

g, h : polynomial.

Remark: We have approximations and recurrence formulae from rational generating functions, however it is difficult (or very tedious task) to obtain explicit formulae from them in general.

Known results for runs

For $x \in \{0, 1\}^n$, let

- (i) $E_{n,m}(x)$, the number of 0^m of size exactly m in x [5], (explicit formulae are not known except for Markov imbedding method.)
- (ii) $G_{n,m}(x)$, the number of 0^m of size greater than or equal to m in x [5, 3],
- (iii) $N_{n,m}(x)$, the number of non-overlapping 0^m in x [6, 9, 14, 3, 5],
- (iv) $M_{n,m}(x)$, the number of overlapping 0^m in x [11, 3, 10, 5, 7]
- (v) $L_n(x)$, the size of the longest run of 0s in x [12, 16, 3, 5],
- (vi) $T_k(x)$, the stopping time t such that 0^k first appear in $x = x_1 \cdots x_t$ [2, 15, 19], and
- (vii) $N_{n,m,\mu}(x)$, the enumeration of 0^m such that we allow μ -letters overlapping with the previous 0^m in the string x [1, 8, 13]. (explicit formulae are not known.)

Example: $E_{8,2} = 1$, $G_{8,2} = 2$, $N_{8,2} = 3$, $M_{8,2} = 4$, and $L_8 = 4$ for
run 00 and 10000100.

Main theorem 1: Distributions of nonoverlapping words

Theorem (Takahashi [21, 20])

Let $X_1^n := X_1 X_2 \cdots X_n$ be finite valued i.i.d. random variables. Let w_1, \dots, w_l be a set of nonoverlapping words. Let $P(w_i)$ be the probability of w_i (e.g. $P(10) = P(1)P(0)$) for $i = 1, \dots, l$. Let

$$A(k_1, \dots, k_l) = \binom{n - \sum_i |w_i| k_i + \sum_i k_i}{k_1, \dots, k_l} \prod_{i=1}^l P^{k_i}(w_i),$$

$$B(k_1, \dots, k_l) = P(\mathbf{N}(w_1, \dots, w_l; X_1^n) = (k_1, \dots, k_l)),$$

$$F_A(z_1, \dots, z_l) = \sum_{k_1, \dots, k_l} A(k_1, \dots, k_l) z^{k_1} \cdots z^{k_l}, \text{ and}$$

$$F_B(z_1, \dots, z_l) = \sum_{k_1, \dots, k_l} B(k_1, \dots, k_l) z^{k_1} \cdots z^{k_l}.$$

Main theorem 1: Distributions of nonoverlapping words

Then

$$A(k_1, \dots, k_l) = \sum B(t_1, \dots, t_l) \binom{t_1}{k_1} \cdots \binom{t_l}{k_l},$$

$$F_A(z_1, z_2, \dots, z_l) = F_B(z_1 + 1, z_2 + 1, \dots, z_l + 1), \text{ and}$$

$$P(N(w_1, \dots, w_l; X_1^n) = (s_1, \dots, s_l))$$

$$= \sum_{\substack{k_1, \dots, k_l: \\ s_1 \leq k_1, \dots, s_l \leq k_l \\ \sum_i |w_i| k_i \leq n}} (-1)^{\sum_i k_i - s_i} \binom{n - \sum_i |w_i| k_i + \sum_i k_i}{s_1, \dots, s_l, k_1 - s_1, \dots, k_l - s_l} \prod_{i=1}^l P^{k_i}(w_i).$$

Here $N(w_1, \dots, w_l; X_1^n) = (s_1, \dots, s_l) \Leftrightarrow w_i$ appear s_i times in X_1^n for all i .

Notations

$$\mathbf{N}'(w_1, \dots, w_l; X_1^n) := (s_1 - s_2, s_2 - s_3, \dots, s_l)$$

if $\mathbf{N}(w_1, \dots, w_l; X_1^n) = (s_1, s_2, \dots, s_l)$.

Example:

$$\mathbf{N}(100, 1000; 10010001) = (2, 1) \text{ and } \mathbf{N}'(100, 1000; 10010001) = (1, 1).$$

Main theorem 2: Distributions of runs

Theorem

Let X_1, X_2, \dots , be i.i.d. binary random variables from P . Let $w_1 \sqsubset w_2 \cdots \sqsubset w_l$ be an increasing sequence of non-overlapping words. Let

$$A(k_1, \dots, k_l) := \binom{n - \sum_i |w_i| k_i + \sum_i k_i}{k_1, \dots, k_l} \prod_{i=1}^l P^{k_i}(w_i),$$

$$B(k_1, \dots, k_l) := P(\mathbf{N}'(w_1, \dots, w_l; X_1^n) = (k_1, k_2, \dots, k_l)),$$

$$F_A(z_1, \dots, z_l) := \sum_{\substack{k_1, \dots, k_l: \\ \sum_i |w_i| k_i \leq n}} A(k_1, \dots, k_l) z^{k_1} \cdots z^{k_l}, \text{ and}$$

$$F_B(z_1, \dots, z_l) := \sum_{\substack{k_1, \dots, k_l: \\ \sum_i |w_i| k_i \leq n}} B(k_1, \dots, k_l) z^{k_1} \cdots z^{k_l}.$$

Main theorem 2: Distributions of runs

Theorem (Continue)

Then

$$F_A(z_1, \dots, z_l) = F_B(z_1 + 1, z_1 + z_2 + 1, \dots, \sum_i z_i + 1) \text{ and}$$

$$P(C_{n,(w_1,\dots,w_l)}(X_1^n) = t) = \sum_{\substack{r, k_1, \dots, k_l: \\ \sum |w_i| k_i \leq n, r \leq \sum k_i \\ t = \sum ik_i - r}} (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \dots, k_l} \\ \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i), \text{ where}$$

$$C_{n,(w_1,\dots,w_l)}(x) := t \text{ if } \sum ik_i = t \text{ and } \mathbf{N}'(w_1, \dots, w_l; x_1^n) = (k_1, k_2, \dots, k_l).$$

Notation

Let $m_1 < \dots < m_l$. For $x \in \{0, 1\}^n$, let

$$D_{n,(m_1,\dots,m_l)}(x) := t \Leftrightarrow \sum ik_i = t \text{ and } \mathbf{N}'(10^{m_1} \dots, 10^{m_l}; 1x) = (k_1, k_2, \dots, k_l),$$

where $1x$ is the concatenation of 1 and x .

C_n does not count runs that start from the head of sample x but D_n does.

Example: $C_{n,100}(00100) = 1$ and $D_{n,2}(00100) = 2$.

Corollary

Let X_1, X_2, \dots, X_n be i.i.d. binary random variables. Let $m_1 < \dots < m_l$ and $w_i = 10^{m_i}$ for $1 \leq i \leq l$. For all $t \geq 0$,

$$\begin{aligned} P(D_{n,(m_1,\dots,m_l)}(X_1^n) = t) \\ = (P(C_{n+1,(w_1,\dots,w_l)}(X_1^{n+1}) = t) - P(0)P(C_{n,(w_1,\dots,w_l)}(X_1^n) = t))/P(1). \end{aligned}$$

Proof) Observe that

$$\begin{aligned} \{0x \mid C_{n+1,(w_1,\dots,w_l)}(0x) = t, |x| = n\} &= \{0x \mid C_{n,(w_1,\dots,w_l)}(x) = t, |x| = n\}, \\ \{1x \mid C_{n+1,(w_1,\dots,w_l)}(1x) = t, |x| = n\} &= \{1x \mid D_{n,(w_1,\dots,w_l)}(x) = t, |x| = n\}. \end{aligned}$$

We have

$$\begin{aligned} P(C_{n+1,(w_1,\dots,w_l)}(X_1^{n+1}) = t) \\ = P(C_{n+1,(w_1,\dots,w_l)}(0X_1^n) = t) + P(C_{n+1,(w_1,\dots,w_l)}(1X_1^n) = t) \\ = P(0)P(C_{n,(w_1,\dots,w_l)}(X_1^n) = t) + P(1)P(D_{n,(w_1,\dots,w_l)}(X_1^n) = t). \end{aligned}$$

Main Theorem2: Distributions of runs

Theorem

Let X_1, X_2, \dots , be i.i.d. binary random variables from $P(X_i = 1) = q$ and $P(X_i = 0) = p$ for all i , where $p + q = 1$.

$$\begin{aligned} 1. P(D_{n, (m_1, \dots, m_l)}(X_1^n) = t) = & \\ & \sum_{\substack{r, k_1, \dots, k_l: \\ \sum (m_i + 1)k_i \leq n + 1, r \leq \sum k_i \\ t = \sum ik_i - r}} (-1)^r \binom{n + 1 - \sum m_i k_i}{k_1, \dots, k_l} \binom{\sum k_i}{r} q^{l-1} p^{\sum m_i} \\ - & \sum_{\substack{r, k_1, \dots, k_l: \\ \sum (m_i + 1)k_i \leq n, r \leq \sum k_i \\ t = \sum ik_i - r}} (-1)^r \binom{n - \sum m_i k_i}{k_1, \dots, k_l} \binom{\sum k_i}{r} q^{l-1} p^{1 + \sum m_i}. \end{aligned}$$

Theorem (Continue)

2. $P(N_{n,m,\mu}(X_1^n) = t) = P(D_{n,(m_1,\dots,m_T)}(X_1^n) = t)$ for all $0 \leq \mu \leq m - 1$,
where $T = \lfloor \frac{n - \mu}{m - \mu} \rfloor$ and $m_i = m - \mu(i - 1)$ for $1 \leq i \leq T$.

$$\begin{aligned} 3. P(G_{n,m}(X_1^n) = t) &= \sum_{t \leq k \leq \lfloor \frac{n+1}{m+1} \rfloor} (-1)^{k-t} \binom{n+1-mk}{t, k-t} q^{k-1} p^{km} \\ &\quad - \sum_{t \leq k \leq \lfloor \frac{n}{m+1} \rfloor} (-1)^{k-t} \binom{n-mk}{t, k-t} q^{k-1} p^{km+1}. \end{aligned}$$

Theorem (Continue)

$$4. P(T_m > n) = P(L_n < m) = \sum_{0 \leq k \leq \lfloor \frac{n+1}{m+1} \rfloor} (-1)^k \binom{n+1-mk}{k} q^{k-1} p^{km} \\ - \sum_{0 \leq k \leq \lfloor \frac{n}{m+1} \rfloor} (-1)^k \binom{n-mk}{k} q^{k-1} p^{km+1}.$$

$$5. P(E_{n,m}(X_1^n) = t) =$$

$$\sum_{\substack{k_1, k_2: \\ (m+1)k_1 + (m+2)k_2 \leq n+1, \\ t \leq k_1 + k_2}} (-1)^{k_1-t} \binom{n+1-mk_1-(m+1)k_2}{k_1, k_2} \binom{k_1+k_2}{t} q^{k_1+k_2-1} p^{k_1m+k_2(m+1)} \\ - \sum_{\substack{k_1, k_2: \\ (m+1)k_1 + (m+2)k_2 \leq n, \\ t \leq k_1 + k_2}} (-1)^{k_1-t} \binom{n-mk_1-(m+1)k_2}{k_1, k_2} \binom{k_1+k_2}{t} q^{k_1+k_2-1} p^{k_1m+k_2(m+1)+1}.$$

Algorithm and Complexity for Computing $P(C_n = t)$

Let $\mathbb{Z}_0 = \{0, 1, 2, \dots\}$ and

$$G_t(n, l, |w_1|, \dots, |w_l|) :=$$

$$\{(r, k_1, \dots, k_l) \in \mathbb{Z}_0^{l+1} \mid \sum_{1 \leq i \leq l} k_i |w_i| \leq n, r \leq \sum_{i=1}^l k_i, \sum_{i=1}^l i k_i - r = t\}.$$

$$P(C_{n, (w_1, \dots, w_l)}(X_1^n) = t) = \sum_{(r, k_1, \dots, k_l) \in G_t} (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \dots, k_l} \\ \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i).$$

$$G(n, l, |w_1|, \dots, |w_l|) :=$$

$$\{(r, k_1, \dots, k_l) \in \mathbb{Z}_0^{l+1} \mid \sum_{1 \leq i \leq l} k_i |w_i| \leq n, r \leq \sum_{i=1}^l k_i, 0 \leq \sum_{i=1}^l i k_i - r \leq t\}.$$

Algorithm and Complexity for computing $P(C_n = h)$, $h = 0, \dots, t$ (Application of Bucket sort algorithm)

Algorithm

1. Initialize $P(C_n = h) = 0$ for all $h = 0, \dots, t$.
2. Enumerate all nonnegative $(r, k_1, \dots, k_l) \in G(n, l, |w_1|, \dots, |w_l|)$.
For each vector $(r, k_1, \dots, k_l) \in G(n, l, |w_1|, \dots, |w_l|)$, set

$P(C_{n, (w_1, \dots, w_l)}(X_1^n) = h) := P(C_{n, (w_1, \dots, w_l)}(X_1^n) = h) + f(r, k_1, \dots, k_l)$, where

$$h = \sum_{i=1}^l i k_i - r \text{ and } f(r, k_1, \dots, k_l) = (-1)^r \binom{n - \sum |w_i| k_i + \sum k_i}{k_1, \dots, k_l} \\ \times \binom{\sum k_i}{r} \prod P^{k_i}(w_i).$$

3. Output $P(C_n = h)$ for all $h = 0, \dots, t$.

Lemma

Assume that $1 \leq |w_1| \leq \dots \leq |w_l|$. Then

$$|\{(k_1, \dots, k_l) \in \mathbb{Z}_0^l \mid \sum_{1 \leq i \leq l} k_i |w_i| \leq n\}| \leq \frac{(n + \sum_i |w_i|)^l}{l! \prod_i |w_i|}.$$

Theorem

Let n be the sample size. For given n, l, t , and w_1, \dots, w_l ,

- 1.

$$|G(n, l, |w_1|, \dots, |w_l|)| \leq (t+1) \left(t + \frac{l(l-1)}{2}\right)^{l-1} ((l-1)!)^{-2} \left(\frac{n}{|w_1|} + 1\right),$$

2. Fix l and t . Then $|G(n, l, |w_1|, \dots, |w_l|)| = O(n)$,

3. Fix l . Assume that $t = O(n^{\frac{\alpha}{l}})$ and $\alpha > 0$. Then

$$|G(n, l, |w_1|, \dots, |w_l|)| = O(n^{\alpha+1}).$$

Theorem (Continue)

4. Let α, β, γ be positive constants, $t \sim \alpha l$, and $l \sim \frac{\log n + \beta}{\gamma}$. Then

$$|G(n, l, |w_1|, \dots, |w_l|)| = O(n^{(2 - \log 2)/\gamma + 1}).$$

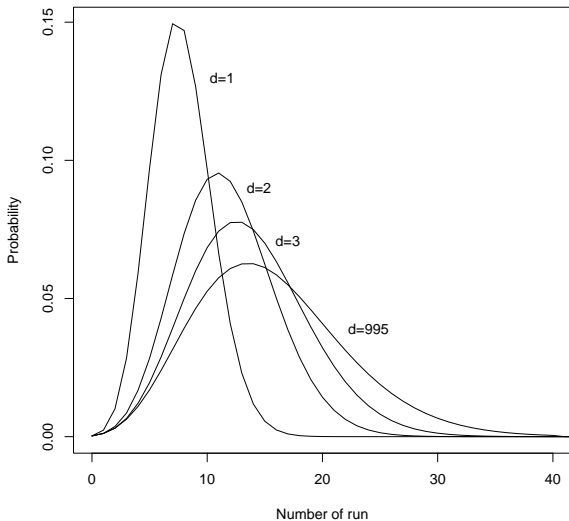
If $\gamma = -\log p$ and $\beta = (r + 1) \log 2 + \log p$ then

$$|P(D_{n, (m_1, \dots, m_l)}(X_1^n) = t) - P(D_{n, (m_1, \dots, m_T)}(X_1^n) = t)| \leq 2^{-r}$$

for all $l < T$, $r > 0$, and t .

If $\gamma = -\log p$, $p = 0.5$ then $(2 - \log 2)/\gamma + 1 = 2.88539$.

Graph of $P(D_{n,(m_1,\dots,m_d)})$ for $d = 1, 2, 3, 995$, $m_1 = 5, m_2 = 6, \dots$, sample size 1000, $P(0)=0.5$.



Reference I

- [1] S. Aki and K. Hirano.
Numbers of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order k .
Ann. Inst. Statist. Math., 52(4):767–777, 2000.
- [2] S. Aki, H. Kuboki, and K. Hirano.
On discrete distributions of order k .
Ann. Inst. Statist. Math., 36:431–440, 1984.
- [3] D. L. Antzoulakos and S. Chadjiconstantindis.
Distributions of numbers of success runs of fixed length in Markov dependent trials.
Ann. Inst. Statist. Math., 53(3):599–619, 2001.
- [4] F. Bassino, J. Clément, and P. Micodème.
Counting occurrences for a finite set of words: combinatorial methods.
ACM Trans. Algorithms., 9(4):Article No. 31, 2010.
- [5] J. C. Fu and M. V. Koutras.
Distribution theory of runs: a Markov chain approach.
J. Amer. Statist. Assoc., 89(427):1050–1058, 1994.

Reference II

- [6] A. P. Godbole.
Specific formulae for some success run distributions.
Statist. Probab. Lett., 10:119–124, 1990.
- [7] A. P. Godbole.
The exact and asymptotic distribution of overlapping success runs.
Comm. Statist. Theory Methods, 21:953–967, 1992.
- [8] S. Han and S. Aki.
A unified approach to binomial-type distributions of order k .
Commun. Statist. Theor. Meth., 29:1929–1943, 2000.
- [9] K. Hirano.
Some properties of the distributions of order k .
pages 43–53, 1986.
Fibonacci Numbers and their Applications, A. N. Phillipou, A. F. Horadam and G. E. Bergum eds, Reidel.
- [10] M. V. Koutras and V. A. Alexandrou.
Non-parametric randomness tests based on success runs of fixed length.
Statist. Probab. Lett., 32:393–404, 1997.

Reference III

- [11] K. D. Ling.
On binomial distributions of order k .
Statist. Probab. Letters, 6:247–250, 1988.
- [12] F. S. Makri, A. N. Philippou, and Z. M. Psillakis.
Shortest and longest length of success runs in binary sequences.
J. Statist. Plan. Inference, 137:2226–2239, 2007.
- [13] F. S. Makri and Z. M. Psillakis.
On l -overlapping runs of ones of length k in sequences of independent binary random variables.
Commun. Statist. Theor. Meth., 44:3865–3884, 2015.
- [14] M. Muselli.
Simple expressions for success run distributions in Bernoulli trials.
Statist. Probab. Lett., 31:121–128, 1996.
- [15] A. N. Philippou, C. Georghiou, and G. N. Philippou.
A generalized geometric distribution and some of its properties.
Statist. Probab. Letters, 1:171–175, 1983.

Reference IV

- [16] A. N. Phillipou and F. S. Makri.
Successes, runs and longest runs.
Statist. Probab. Lett., 4:211–215, 1986.
- [17] M. Régnier and W. Szpankowski.
On pattern frequency occurrences in a Markovian sequence.
Algorithmica, 22(4):631–649, 1998.
- [18] S. Robin and J. J. Daudin.
Exact distribution of word occurrences in a random sequence of letters.
J. Appl. Probab., 36(1):179–193, 1999.
- [19] V. R. R.Uppuluri and S. A. Patil.
Waiting times and generalized Fibonacci sequences.
Fibonacci Quart., 21:242–249, 1983.
- [20] H. Takahashi.
The explicit formulae for the distributions of nonoverlapping words and its applications to statistical tests for pseudo random numbers, May 2021.
[Arxiv 2105.05172](https://arxiv.org/abs/2105.05172).

Reference V

[21] H. Takahashi.

The explicit formula for the distributions of nonoverlapping words.
IEICE Technical Report IT2021-123, 121(428):234–236, Mar 2022.