Accurate computation of the variance of the number of missing words in a random string

Paul Leopardi

Thanks: Jörg Arndt, Richard Brent, Sylvain Forêt, Judy-anne Osborn.

Mathematical Sciences Institute, Australian National University.

For presentation at 4th International Conference on Combinatorial Mathematics and Combinatorial Computing, Auckland New Zealand, December 2008.

AUSTRALIAN RESEARCH COUNCIL
Centre of Excellence for Mathematics and Statistics of Complex Systems

ARC

Topics

- Overlapping serial tests
- Analysis of the problem: missing words in a string
- Word overlap correlations
- Enumeration of correlations (and generating functions)
- Open problems

Overlapping serial tests (1993)

Overlapping serial tests (Marsaglia and Zaman, 1993) use an alphabet of size α, form a pseudorandom string of length $N = 2^{21}$, and examine the overlapping words of length T.

Number of missing words should be approximately normal with mean μ and variance σ^2:

Test	α	T	μ	σ
OPSO:	1024	2	141909.4653	290.27
OQSO:	32	4	141909.4737	290
DNA:	4	10	141910.5378	290

Overlapping serial tests (1995)

(Marsaglia, 1995) has the revised values:

Test	α	T	μ	σ	(σ was)
OPSO:	1024	2	141909.60	290.46	290.27
OQSO:	32	4	141909.4737	295	290
DNA:	4	10	141910.5378	339	290

OQSO: "I don't know, and doubt that I ever will know, the true variance. There are just too many kinds of pairs of 4-letter words to undertake finding all the necessary generating functions."

DNA: "It appears a formidable task to find the exact variance for the DNA test."

Overlapping serial tests (2008)

Values calculated using (Noonan, Zeilberger, 1999), (Rivals, Rahmann, 2003) and (Rahmann, Rivals, 2003):

Test	α	T	μ	σ
OPSO:	2^{10}	2	141909.3299550069	290.4622634038
OQSO:	32	4	141909.6005321316	294.6558723658
DNA:	4	10	141910.4026047629	337.2901506904

- Calculation of σ for OPSO uses **6** generating functions;
- OQSO uses **55**; DNA uses **4592**.

Strings, words, indicator variable

We analyze the problem: find the distribution of the number of missing words in a random string.

Alphabet size is α, equally likely.

String length is N. Word length is T.

Words overlap. The string S contains $N - T + 1$ words.

There are α^N possible strings S_i, α^T possible words W_j.

Define indicator $v_{i,j} := 1 \Leftrightarrow$ word W_j is missing from string S_i.

Number of missing words X

The number of words missing from string S_i is

$$X_i := \sum_j v_{i,j}.$$

X is the number of words missing from a random string S.

For constant $\lambda := N/\alpha^T$ as $N \to \infty$,
X is asymptotically normal. (Rukhin 2002)

Pair absence probability, generating functions

The probability that both words W_j and W_k are missing from a random string S is

$$a_{j,k} := \alpha^{-N} \sum_i v_{i,j} v_{i,k}.$$

Generating functions:

$$A_{j,k} : [z^N]\, A_{j,k}(z) = a_{j,k},$$
$$A_j : [z^N]\, A_j(z) = a_{j,j}.$$

Expected value, variance

The expected value of X is

$$\mathbf{E}[X] = \alpha^{-N} \sum_i X_i = \alpha^{-N} \sum_i \sum_j v_{i,j}$$
$$= \sum_j a_{j,j}.$$

The variance is $\mathbf{var}[X] = \mathbf{E}[X^2 - X] - \mathbf{E}[X] - \mathbf{E}[X]^2$, with

$$\mathbf{E}[X^2 - X] = \alpha^{-N} \sum_i \sum_{j \neq k} v_{i,j} v_{i,k}$$
$$= \sum_{j \neq k} a_{j,k}.$$

Word overlap correlation vectors

Words B, C of length T, $B_0 \dots B_{T-1}$ etc.

(Word overlap) correlation vector BC:
$BC_s = 1 \Leftrightarrow B_{r+s} = C_r\,,\ r = 0 \dots T - S - 1\,.$

B :	D	A	N	G	E	R	
C :	A	N	G	E	R	S	
		A	N	G	E	R	S
			...				
BC :	0	1	0	0	0	0	

Correlation vectors BB, CC are called autocorrelations.

(Guibas and Odlyzko 1981; Rivals and Rahmann 2003)

Correlation polynomials

For correlation vector v, the correlation polynomial P_v is

$$P_v(z) := v_0 + v_1 z + \ldots + v_{T-1} z^{T-1}.$$

For $P_j := P_{W_j W_j}$, the generating function A_j is

$$A_j(z) = \frac{P_j(z/\alpha)}{(z/\alpha)^T + (1-z)P_j(z/\alpha)}.$$

(Guibas and Odlyzko 1981; Rahmann and Rivals 2003, Lemma 2.1)

Correlation matrices and correlation classes

For $P_{j,k} := P_{W_j,W_k}$ etc. the correlation matrix is

$$M_{j,k}(z) := \begin{bmatrix} P_{j,j}(z) & P_{j,k}(z) \\ P_{k,j}(z) & P_{k,k}(z) \end{bmatrix}.$$

Given $M := \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$ define $M^V := \begin{bmatrix} m_{22} & m_{21} \\ m_{12} & m_{11} \end{bmatrix}$,

$R(M) := m_{11} + m_{22} - m_{12} - m_{21}.$

Define the equivalence class $[M] := \{M, M^T, M^V, M^{TV}\}$, so

$$[M_{j,k}(z) = M_{j,k}(z), M^T_{j,k}(z), M_{k,j}(z), M^T_{k,j}(z)\}.$$

Note $M' \in [M] \Rightarrow \det M' = \det M$ and $R(M') = R(M)$.

(Rahmann and Rivals 2003, Lemma 3.2)

Generating function for pairs of words

For $Q_{j,k}(z) := \det M_{j,k}(z), \quad R_{j,k}(z) := R(M_{j,k}(z))$, the generating function $A_{j,k}$ for the pair W_j, W_k is given by

$$A_{j,k}(z) = \frac{Q_{j,k}(z/\alpha)}{(1-z)Q_{j,k}(z/\alpha) + (z/\alpha)^T R_{j,k}(z/\alpha)}.$$

(Rahmann and Rivals 2003, Lemma 3.2)

Also (Goulden and Jackson 1979, 1983; Guibas and Odlyzko 1981; Noonan and Zeilberger 1997; Rukhin 2002).

Set partitions, restricted growth strings

We could simply sum $a_{j,k}$ for all $\alpha^{2T} - \alpha^{T}$ word pairs $W_j \neq W_k$, but for Marsaglia's tests, $\alpha^{2T} = 2^{40}$.
So instead we enumerate correlation classes and count the word pairs for each class.

Word pairs W_j, W_k with β different letters
$\longrightarrow$ partition of $\{0, \dots, 2T-1\}$ into β nonempty subsets
$\longleftrightarrow$ restricted growth string of length $2T$ with β different letters.

S is a restricted growth string if $S_k \leqslant S_j + 1$ for each j from 0 to $k-1$, for k from 1 to $2T-1$.

Set partitions, restricted growth strings

Each permutation of the alphabet preserves the correlation matrix. The set of word pairs with β different letters splits into orbits under $\mathbb{S}_\alpha$ of size

$$\frac{\alpha!}{(\alpha-\beta)!}.$$

The number of partitions of $\{0,\ldots,2T-1\}$ into exactly β nonempty subsets is the second kind Stirling number $S(2T,\beta)$.

If $\alpha \leqslant 2T$, the total number of word pairs is

$$\alpha^{2T} = \sum_{\beta=1}^{\alpha} \frac{\alpha!}{(\alpha-\beta)!} S(2T,\beta).$$

Enumeration by set partitions

Define $n[M](\alpha) = \sharp\{(j,k) \mid M_{j,k} = [M]\}$,
the number of word pairs for correlation class $[M]$.

For $\alpha \leqslant 2T$, to determine all correlation classes $[M]$,
and find $n[M](\alpha)$ for each,

Keep a count for each correlation class encountered so far;
For each β from 1 to α:

- For each restricted growth string of length $2T$ with exactly β different letters:
  1. Find the correlation class for the corresponding word pair;
  2. Add $\frac{\alpha!}{(\alpha-\beta)!}$ to the count for the class.

Population of each correlation class

For each correlation class $[M]$, $n[M](\alpha)$ is a polynomial in α of maximum degree $2T$.

For $\alpha > 2T$, to find $n[M](\alpha)$, first find $n[M](\gamma)$ for γ from 1 to $2T$ and interpolate the polynomial.

In the case of Marsaglia's tests:

Test	α	T	Classes	Method
OPSO:	1024	2	6	(Rahmann and Rivals 2003)
OQSO:	32	4	55	Polynomial interpolation
DNA:	4	10	4592	Exhaustive enumeration

Number of correlation classes

Define $b(T, \alpha)$ to be the number of correlation classes for unequal strings of length T and alphabet size α.

The set of classes remains unchanged for $\alpha > 2T$.

The number of classes $b(T, \alpha)$ for small T is:

α	1	2	3	4	5	6	7	8	9	10
2	1	3	11	31	87	193	415	839	1632	3004
3	1	6	20	54	141	322	655	1322	2506	4577
4	1	6	20	55	141	324	657	1329	2515	4592
$2T$	1	6	20	55	141	324	657	????	????	????

See A152139, A152959, Online Encyclopedia of Integer Sequences.

Some open problems

1. "Characterize and efficiently enumerate $\mathbf{2} \times \mathbf{2}$, and more generally, $\boldsymbol{k} \times \boldsymbol{k}$ matrices of correlation vectors between $\boldsymbol{k}$ pairwise different [words], and find the number of such matrices.
   Compute the number of $\boldsymbol{k}$-tuples of words that share a given correlation matrix."
   (Rahmann and Rivals 2003)

2. For $\boldsymbol{T} > \mathbf{2}$, $\boldsymbol{\lambda} := \boldsymbol{N}/\boldsymbol{\alpha}^{\boldsymbol{T}}$ constant as $\boldsymbol{N} \to \infty$, find a high order asymptotic expansion for $\mathbf{var}[\boldsymbol{X}]$.
   (Rukhin 2002; Rahmann and Rivals 2003)

3. Does $\boldsymbol{b}(\boldsymbol{T}, 4) = \boldsymbol{b}(\boldsymbol{T}, \mathbf{2}\boldsymbol{T})$ for all $\boldsymbol{T}$?