Parabola Volume 61, Issue 1 (2025)

An investigation of coverage of digits by natural sequence

Sin Keong Tong[1]

1 Introduction

In the decimal expansion of the first billion digits of π, the 8-digit birthdate '22121887' of the Indian mathematician Srinivasa Ramanujan is found 9 times with gaps ranging from 11,872,699 to 442,965,910. On the premise that π is a universal[2] number, it is expected that all 8-digit numbers would eventually appear among the digits of π.

For each positive integer k, let $T_k = \{\overbrace{0\ldots0}^{k}, \ldots, \overbrace{9\ldots9}^{k}\}$ denote the set of all 10^k sequences of k decimal digits. The first 33 consecutive digits of π *covers* T_1 in the sense that each of the numbers in $T_1 = \{1, 2, \ldots, 9\}$ appear among these 33 digits. The first occurrence of the 10 digits is highlighted in red:

314159265358979323846264338327950

The set of 2-digit sequences T_2 is covered by the first consecutive 606 digits of π:

3141592653589793238462643383279502884197169399375105820974944592307 8
16406286208998628034825342117067982148086513282306647093844609550582
23172535940812848111745028410270193852110555964462294895493038196442
88109756659334461284756482337867831652712019091456485669234603486104
54326648213393607260249141273724587006606315588174881520920962829254
09171536436789259036001133053054882046652138414695194151160943305727
03657595919530921861173819326117931051185480744623799627495673518857
52724891227938183011949129833673362440656643086021394946395224737190
702179860943702770539217176293176752384674818467669405132000568

The number 68 completes the coverage of T_2.

The *completion length* for T_k is the length of a sequence of digits of π that covers T_k. This length depends strongly of the starting position of the sequence. Starting at the first digit of π, the completion lengths for T_1 and T_2 are as above: 33 and 606. However, by starting at 10 random positions in the digits of π, I calculated the completion lengths to be 10, 11, 18, 23, 26, 32, 32, 34, 49, 51 for T_1 and 349, 362, 381, 401, 494, 496, 506, 516, 568, 839 for T_2, all distinct and dissimilar lengths.

This article develops three methods of calculating the expected distribution of the completion length for T_k. The mean, mode and median of the distribution are then compared with the digits of classical and computer generated integer sequences. It will also be demonstrated that there are uncountable infinite independent randomness tests that one can devise for any infinite sequence.

[1]Sin Keong Tong is a Problem Editor for *Parabola*.

[2]A number is said to be *universal* if all finite sequence appear in its decimal expansion.

2 The combinatorics method

This section illustrates by way of an example a method for calculating the distribution of the completion lengths of T_k.

Example 1. How many 12-digit integer sequences cover T_1?

To find this number, note that there are two ways to partition 12 as sum of 10 natural numbers.

(i) $12 = 3 + 1 + \cdots + 1$. The number of such partitions is $\binom{10}{1}\binom{9}{9} = 10$.
The number of permutations of each partition is $\frac{12!}{3!1!^9} = 79,833,600$.
Thus, there are $10 \times 79833600 = 798,336,000$ 12-digit numbers.

(ii) $6 = 2 + 2 + 1 + \cdots + 1$. The number of such partitions is $\binom{10}{2}\binom{8}{8} = 45$.
The permutation of each partition is $\frac{12!}{2!^2 1!^1 0} = 119,750,400$.
Thus, there are $45 \times 119750400 = 5,388,768,000$ 12-digit numbers.

The number of 12-digit numbers that cover T_1 is

$$798336000 + 5388768000 = 6,187,104,000\,.$$

This is approximately 0.006 of all 12-digit numbers. This method is not feasible for larger values of k. The number of partitions for $n = 80$ as sum of 10 positive integers is 15,796,476. A Google search indicates that there is no agreed-on most efficient method for generating these partitions.

3 Concepts and notation

For this article, n elements are chosen from T_k to form a string s. For the moment, it is convenient to think of s as consecutive k digits. In this case, the length of s is $n - k + 1$. The extension to building a string with non-overlapping elements of k digits is discussed in Section 6.4.

Definition 2. The *pattern set* $C(n, m)$ is the collection of pattern elements $s = \{a_1 \ldots a_n\}$ where $a_1, \ldots, a_n$ are each any of m elements.

In the pattern set, the number of occurrences of an element and the permutation between the elements are important. For instance, the element aab represents also bba.

Example 3. Consider the patterns created by painting $n = 4$ shapes in a straight line by exactly $m = 1, 2, 3, 4, 5$ patterns:

$$\begin{aligned}
&C(4,1) = \{aaaa\}\\
&C(4,2) = \{aaab, aaba, abaa, aabb, abab, abba, abbb\}\\
&C(4,3) = \{aabc, abac, abca, abbc, abcb, abcc\}\\
&C(4,4) = \{abcd\}\\
&P(4,5) = \{\emptyset\}\,.
\end{aligned}$$

Definition 4. The *configuration set* $P(n, m, r)$ is extension of $C(n, m)$ where the elements are chosen from a set $P = \{p_1, \ldots, p_r\}$.

Example 5. The configurations created by painting of 3 distinguishable houses with m patterns along a straight line by three paints p, q, r are:

$P(3, 1, 3) = \{ppp, qqq, rrr\}$
$P(3, 2, 3) = \{ppq, ppr, pqp, pqq, prp, prr, qpp, qpq, qqp, qqr, qqr, qrq, rpp, rpr, rqq, rqr, rrp, rrq\}$
$P(3, 3, 3) = \{pqr, prq, qpr, qrp, rpq, rqp\}$
$P(3, 4, 3) = \{\emptyset\}$.

Let $P_k(n, m)$ briefly denote the set $P(n, m, 10^k)$. This is the set that will be of interest when considering the set T_k.

The separation of $C(n, m)$ from $P(n, m)$ and $P_k(n, m)$ permits the recursive method in Section 4 to be also applied to related problems.

4 Recursive method

In the example below, assume that a, b, c, d are distinct symbols.

Example 6.
For $n = 1$, $C(1, 1) = \{a\}$. Therefore, $|P_1(1, 1)| = 1 \times 10 = 10$.
For $n = 2$, $C(2, 1) = \{aa\}$. Therefore, $|P_1(2, 1)| = 1 \times 10 = 10$.
$C(2, 2) = \{ab\}$. Therefore, $|P_1(2, 2)| = 1 \times 10 \times 9 = 90$.
For $n = 3$, $C(3, 1) = \{aaa\}$. Therefore, $|P_1(3, 1)| = 1 \times 10 = 10$.
$C(3, 2) = \{aab, aba, abb\}$. Therfore, $|P_1(3, 1)| = 3 \times 10 \times 9 = 270$.
$C(3, 3) = \{abc\}$. Therefore, $|P_1(3, 1)| = 1 \times 10 \times 9 \times 8 = 720$.
For $n = 4$, $C(4, 1) = \{aaaa\}$. Therefore, $|P_1(4, 1)| = 1 \times 10$.
$C(4, 2) = \{aaab, aaba, aabb, abbb, aabb, abab, abba\}$.
Therefore, $|P_1(4, 2)| = 7 \times 10 \times 9 = 630$.
$C(4, 3) = \{aabc, abac, abca, abbc, abcb, abcc\}$.
Therefore, $|P_1(4, 3)| = 6 \times 10 \times 9 \times 8 = 4320$.
$C(4, 4) = \{abcd\}$.
Therefore, $|P_1(4, 4)| = 10 \times 9 \times 8 \times 7 = 5040$.

Note that the total configuration count for each of $n = 1, 2, 3, 4$ is 10, 100, 1000, 10000, respectively. By extending the calculations above, the reader should be able to see that

$$10^8 = |C(4, 1)|(100)_1 + |C(4, 2)|(100)_2 + |C(4, 3)|(100)_3 + |C(4.4)|(100)_4 \qquad (1)$$

where $(x)_k = \prod_{i=0}^{k-1}(x - i)$ is the falling factorial.

Lemma 7. *Let k, m, n be positive integers. Then $|C(n, 1)| = 1$,*

$$|C(n, m)| = |C(n-1, m-1)| + m|C(n-1, m)| \quad \textit{and} \quad \sum_{m=1}^{n} (10^k)_m \, |C(n, m)| = 10^{kn} .$$

Proof. $C(n, 1)$ is the pattern $\overbrace{a \dots a}^{n}$. Therefore, $|C(n, 1)| = 1$.

Let $s = a_1 \dots a_{n-1} a_n \in C(n, m)$ cover $D = \{d_1, \dots, d_m\}$, and denote $s' = a_1 \dots a_{n-1}$. If s' covers D, then $s'' \in C(n, m)$ is formed by appending any of the m pattern elements in D to s'; otherwise, s' covers $D \setminus \{d_i\}$, and $a_1 \dots a_{n-1} d_i \in C(n, m)$. It follows that $|C(n, m)| = |C(n - 1, m - 1)| + m|C(n - 1, m)|$.

The remaining sum is a generalisation of Example 6. Let $e = a_1 \dots a_n \in C(n, m)$ where $\{a_1, \dots, a_n\}$ are pattern elements of length k. By definition, e generates $(10^k)_m$ configurations in T_{kn}. Conversely, every configuration in T_{nk} is mapped to a unique pattern in $C(n, m)$. The proof follows by summing $(10^k)_m \, |C(n, m)|$ over m. $\square$

Table 1 demonstrates the application of Lemma 7 (ii) for $1 \leq n \leq 12$.

$n \setminus m$	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0
3	1	3	1	0	0	0	0	0	0	0
4	1	7	6	1	0	0	0	0	0	0
5	1	15	25	10	1	0	0	0	0	0
6	1	31	90	65	15	1	0	0	0	0
7	1	63	301	350	140	21	1	0	0	0
8	1	127	966	1701	1050	266	28	1	0	0
9	1	255	3025	7770	6951	2646	462	36	1	0
10	1	511	9330	34105	42525	22827	5880	750	45	1
11	1	1023	28501	145750	246730	179487	63987	11880	1155	55
12	1	2047	86526	611501	1379400	1323652	627396	159027	22275	1705

Table 1: $|C(n, m)|$

$n \setminus m$	1	2	3	4	5	6	7	8	9	10
1	10	0	0	0	0	0	0	0	0	0
2	10	90	0	0	0	0	0	0	0	0
3	10	270	720	0	0	0	0	0	0	0
4	10	630	4320	5040	0	0	0	0	0	0
5	10	1350	18000	50400	30240	0	0	0	0	0
6	10	2790	64800	327600	453600	151200	0	0	0	0
7	10	5670	216720	1764000	4233600	3175200	604800	0	0	0
8	10	11430	695520	8573040	31752000	40219200	16934400	1814400	0	0
9	10	22950	2178000	39160800	210198240	400075200	279417600	65318400	3628800	0
10	10	45990	6717600	171889200	1285956000	3451442400	3556224000	1360800000	163296000	3628800
11	10	92070	20520720	734580000	7461115200.	27138434400.	38699337600	21555072000	4191264000	199584000
12	10	184230	62298720	3081965040	41713056000	200136182400	379449100800	288538588800	80831520000	**6187104000**

Table 2: $|P_1(n, m)|$

Table 2 is constructed by Lemma 7 (iii). The number $|P_1(12, 10)|$ calculated in Example 1 is highlighted as the blue $P(12, 10)$ in Table 2.

5 A combinatorial expression for $C(n, m)$

This section delivers the following combinatorial expressions for $|C(n, m)|$.

Theorem 8.

$$|C(n, m)| = \sum_{i=1}^{m} \frac{(-1)^{m+i} i^n}{m!(m-i)!} .$$

Two different proofs will be given for this theorem. The first proof is motivated by observations of the patterns in Table 1 and involves solving difference equations. The second proof is obtained by using the Inclusion-Exclusion Principle.

Algebraic proof. By lemma 7 and highlighted by the construction of Table 1, the numbers $C(n, m)$ are determined by the value $|C(1, m)| = 1$ and the recursive identity $|C(n, m)| = |C(n-1, m-1)| + m|C(n-1, m)|$. Define

$$F(n, m) = \sum_{i=1}^{m} \frac{(-1)^{m+i} i^n}{m!(m-i)!} .$$

Note that $F(1, m) = a_{1,1}(1)^1 = \frac{(-1)^{1+1}}{1!(1-1)!} = 1 = |C(n, m)|$. Also, note that

$$\begin{aligned}
F(n-1, m-1) + mF(n-1, m) &= \sum_{i=1}^{m-1} a_{m-1,i}\, i^{n-1} + m \sum_{i=1}^{m} a_{m,i}\, i^{n-1} \\
&= a_{m,m}\, m^n + \sum_{i=1}^{m-1} (a_{m-1,i} + m a_{m,i}) i^{n-1} \\
&= a_{m,m}\, m^n + \sum_{i=1}^{m-1} \left(\frac{(-1)^{m+i-1}}{i!(m-i-1)!} + \frac{m(-1)^{m+i}}{i!(m-i)!} \right) i^{n-1} \\
&= a_{m,m}\, m^n + \sum_{i=1}^{m-1} \frac{(-1)^{m+i-1}\, i^{n-1}}{i!(m-i)!} (m-i-m) \\
&= a_{m,m}\, m^n + \sum_{i=1}^{m-1} \frac{(-1)^{m+i}\, i^n}{i!(m-i)!} \\
&= \sum_{i=1}^{m} \frac{(-1)^{m+i}\, i^n}{(m)!(m-i)!} \\
&= F(n, m) .
\end{aligned}$$

Since $F(n, m)$ and $|C(n, m)|$ satisfy the same recursive identity, $F(n, m) = |C(n, m)|$. $\square$

Proof using the Inclusion/exclusion Principle. By definition, the configuration set $P(n, m)$ is formed by the permutation of m pattern elements for $C(n, m)$. Thus,

$$|P(n, m)| = m!|C(n, m)| . \tag{2}$$

For a pattern string s_p of n items where there are p choices of pattern elements chosen from m elements, the number of pattern strings that can be formed is given by $\binom{m}{p}p^n$. By the Inclusion-Exclusion Principle [1] and Equation (2),

$$|P(n,m)| = \binom{m}{m}m^n - \binom{m}{m-1}(m-1)^n + \cdots + (-1)^{m-1}\binom{m}{1}(1)^n .$$

Therefore, $m!|C(n,m)| = \sum_{i=1}^{m}(-1)^{m-i}\frac{m!}{(m-i)!i!}\, i^n$, and the theorem follows. □

Despite of the elegance of Theorem 8, the instruction sets for computers are not suitable for division, and errors were experienced for $n \geq 40$.

6 Computing analysis

6.1 Density function distribution

Recall that $|P_k(n, 10^k)|$ is the number of strings of length n that cover T_k. In this section, the results obtained for the theoretical calculations in Section 4 are compared with the digit distribution of a number of classical and computer generated sequences.

Define $A_k(n, 10^k)$ and $B_k(n, 10^k)$ as the cumulative and frequency density functions of integer strings of length n that cover T_k:

$$A_k(n, 10^k) = \frac{|P_k(n, 10^k)|}{10^{nk}}$$
$$B_k(n, 10^k) = \frac{A_k(n, 10^k) - A_k(n-1, 10^k)}{10^{nk}} .$$

Figure 1 shows the plot of n against the density function $B_1(n, 10)$ for $n = 10, ..., 80$.

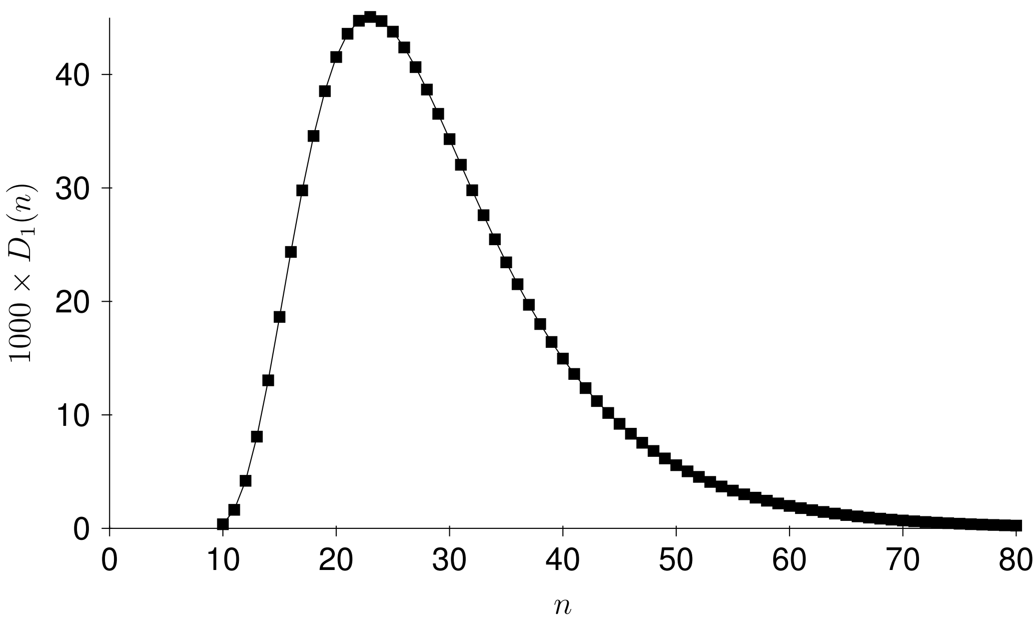

Figure 1: Plot of theoretical values of $B_1(n, 10)$

The number of integer partitions grows very rapidly. The mean, mode, median for $C(n, 10)$ with $1 \leq n \leq 80$ are 29.1, 23 and 26, respectively. It is calculated that $A_1(80, 100) = 0.9978$, indicating that the mean is actually higher than 29.1.

6.2 Dataset analysis for $k = 1$

Next, we turn the attention to the authentication of theoretical results with decimal expansion of classical irrational numbers as well computer generated numbers. Nine datasets are analysed to as the consistency of the expected values of mean, mode and median for $B_1(n, 10)$ and $B_2(n, 100)$:

A dataset of 1 billion digits from the Python random function.
The first 1 billion digits of π.
The first 1 billion digits of $\sqrt{2}$.
A dataset consisting 100,663,295 digit using a VBA Version 1640.
The first 1 billion digits of e.
GIMPS prime $2^{74,207,281} - 1$ with 22,338,618 digits.
The 10,000,000th Fibonacci number with 2,089,877 digits.
A dataset of 1 billion digits from the Matlab random function.
Dataset formed by the concatenation of the numbers $000000, 000001, \ldots, 999999$.

Dataset	Size	Mean	Mode	Median	Max Length	Min Length
Calculated	-	29.1	23	26	80	10
Python	10^9	29.3	23	26	217	10
π	10^9	29.3	23	26	196	10
$\sqrt{2}$	10^9	29.3	23	26	195	10
VBA	150,994,927	29.3	23	26	187	10
e	10^9	29.3	23	26	199	10
$M49$	22,338,589	29.3	23	26	166	10
F_{30000}	1,437,821	29.3	23	26	165	10
Matlab	999,999,964	29.3	23	26	190	10
000000-999999	6,000,000	52.7	55	54	60	19

Table 3: Distribution of covering length of datasets for $k = 1$

6.3 Dataset analysis for $k = 2$

For $k = 2$, the two ways to extract 2 digits are by adjacent digits and non-overlap digits, respectively. Take the first 8 digits '31415926' of π; then the natural way is to extract consecutive digits yielding 31,14,41,15,59,92,26. However, this is problematic because the last digit of a number matches the beginning digit of the next number. On the other hand, the non-overlap extraction yields 31,41,59,26 but may not reflect the intuition of the 2 digits from the string. Table 4 below provides the analysis for both extraction methods. The results for both are almost identical but next section will demonstrate that this is necessarily the case.

Dataset	Median		Mode		Mean		SD	
	adjacent	non-overlap	adjacent	non-overlap	adjacent	non-overlap	adjacent	non-overlap
MS Randomiser	496	496	460	461	518	518	127	126
$\sqrt{2}$	497	497	460	459	519	618	127	125
π	497	497	458	460	519	519	127	126
Fibonacci	497	499	460	460	517	519	125	123
e	495	494	465	465	517	517	126	126
M49 prime	497	497	460	460	518	518	127	126

Table 4: Distribution of covering length of datasets for $k = 2$

6.4 Application of Cantor uncountable infinity to digit extraction

The sequence $s_1 = \overbrace{|00|01|10|11|11|10|01|00|}^{\text{Repeat}} \cdots$ covers the digit pairs $00, 01, 10, 11$ uniformly by non-overlapping extraction. For $(8n + 1)$ single digits in s_1, the frequency counts for $00, 01, 10, 11$ are in the ratio $5 : 3 : 3 : 5$, respectively, for adjacent extraction. s_1 is an example of a sequence where distribution formed by non-overlap extraction is uniform and extraction by adjacent digits fails.

Consider the sequence $s_2 = 0|0110|0110|0110| \cdots$. It is formed by concatenating $\{00, 11\}$ and is missing $\{01, 10\}$. Thus, the non-overlap extraction of s_1 does not cover $\{01, 01, 10, 11\}$ uniformly. However, the first $(4n+1)$ 2-digits adjacent extraction results a frequency count n for each of the pairs $00, 01, 10, 11$. The sequence s_2 is an example where extraction of adjacent digits is uniform but the extraction by non-overlap digit may fail.

The above examples demonstrate that it insufficient to determine the randomness of integer sequences by sampling adjacent digits. Both sequences would not satisfy the distribution for $k = 3$. Take an irrational number $d = d_1 d_2 d_3 \ldots$, where each d_i is a k-digit number. For instance, the extraction for k digit numbers could be from positions $d_1, d_1 + d_2, d_1 + d_2 + d_3, \ldots$. By the construction of Georg Cantor [2], there are uncountable infinity irrational numbers.

A necessary condition for an integer sequence to be random is that every term is independent of the preceding terms. Because of the power series expansions for $\pi, e, \sqrt{2}, F_{1000000}$, this independence cannot be guaranteed, as opposed to that of a prime number. There are uncountable infinity ways of conducting randomness test for the digits of an irrational number. The concept of digits of irrational number such as π are random is therefore not well defined.

7 Conclusion

Three methods of calculating the distribution for $P_k(n, m)$ were presented. The consistency of the distribution between the datasets and expected distribution for $P_k(n, m)$ is rather staggering. It is beyond the power of laptop computers running under Python to perform the necessary calculations for $k \geq 2$. It is reported on the internet that the sequence '0123456789' has not been found in the digits of e. It would of interest to calculate $C(n, 10^{10})$.

8 Acknowledgement

I like to thank Arnaud Brothier, David Angell and Thomas Britz for reviewing the document, adding many insights.

References

[1] R.A. Brualdi, *Introductory Combinatorics* (5th ed.), Prentice-Hall, 2010.

[2] J.W. Dauben, *Georg Cantor - His Mathematics and Philosophy of the Infinite*, Princeton University Press, 1979.