On the Number of Occurrences of the Digit 1 in the Sequence of Positive Integers Less Than n

Curtis Cooper and Robert E. Kennedy
Department of Mathematics, Central Missouri State University
Warrensburg, MO 64093

Introduction. In [1], Trollope discovered a formula for the sum of the digital sums of the positive integers (written in a generalized base) less than n. In [2], he found a particular expression for this sum when the base is two. Here, we are interested in a slightly different problem. Specifically, we will determine a formula for the number of occurrences of the digit 1 in the sequence of positive integers (written in base ten) less than n. Although Trollope's problem and this problem are different, we will proceed in much the same spirit as Trollope.

Initially, our motivation for studying this problem was our interest in digit problems. We had no application in mind when we began this investigation. However, as we worked on this problem, we were able to create a "practical" problem which is related to the number of 1's problem.

> Suppose you own a company which types page numbers on book pages. You would like to obtain estimates for the number of times the number keys on the typewriter are used. In particular,
>
> 1. How much use will the "1" key get when page numbers are typed on an n page book?
> 2. Will the number keys wear evenly? That is, will the "1" key get as much use as the "2" key, etc.?

With this justification of our work, we now set out to answer the number of 1's problem and discuss some related questions.

Notation and Basic Results. We begin with some notation. Let n be a positive integer. Let $D_1(n)$ denote the number of occurrences of the digit 1 in the sequence $1, 2, \ldots, n-1$. For example, $D_1(100) = 20$ since there are twenty 1's in the sequence $1, 2, 3, \ldots 98, 99$. Ten 1's occur in the units digits of $1, 11, 21, \ldots, 91$ and ten 1's occur in the tens digits of $10, 11, 12, \ldots, 19$.

Next, we define a notation which will be used extensively throughout this paper. The idea is due to Kenneth Iverson, the creator of the programming language APL, and is discussed in [3]. Suppose that k is an integer and $P(k)$ is some statement about k which is either true or false. Then

$$[P(k)] = \begin{cases} 1, & \text{P(k) is true;} \\ 0, & \text{P(k) is false.} \end{cases}$$

Our first result is a statement about the number of 1's in the sequence of positive integers up to a digit times a power of ten. The proof of this formula follows from a straightforward counting argument, and will be omitted.

<u>Lemma 1</u>. Let d be a non-zero digit and m a non-negative integer. Then

$$D_1(d \cdot 10^m) = [d > 1] \cdot 10^m + d \cdot m \cdot 10^{m-1}.$$

As an example of Lemma 1, we have that

$$D_1(2000) = [2 > 1] \cdot 10^3 + 2 \cdot 3 \cdot 10^2 = 1600.$$

Next, let n be a positive integer with decimal representation

$$n = \sum_{k=0}^{m} a_k \cdot 10^k.$$

Also, let

$$n_i = \sum_{k=0}^{i} a_k \cdot 10^k, \text{ for } i \geq 0; \quad n_{-1} = 0.$$

For example, $758334109_4 = 34109$. Note that for any $i \geq -1$, $0 \leq n_i < 10^{i+1}$.

Now, we make the important observation, i.e.,

$$\begin{aligned} D_1(n) &= D_1(a_m \cdot 10^m + n_{m-1}) \\ &= D_1(a_m \cdot 10^m) + [a_m = 1] \cdot n_{m-1} + D_1(n_{m-1}). \end{aligned}$$

Then, using mathematical induction on the number of digits in n, the above equation and Lemma 1, we have the following more general result.

<u>Lemma 2</u>.

$$D_1(n) = \sum_{k=0}^{m} a_k \cdot k \cdot 10^{k-1} + \sum_{k=0}^{m} [a_k > 1] \cdot 10^k + \sum_{k=0}^{m} [a_k = 1] \cdot n_{k-1}.$$

From this formula, we can obtain a main term and a big-oh term for $D_1(n)$. To do this, we first note that

$$\sum_{k=0}^{m} a_k \cdot k \cdot 10^{k-1} - \frac{1}{10}nm \leq 9 \sum_{k=0}^{m} (k-m) \cdot 10^{k-1}.$$

Next, by the "perturbation method" ala [3], we have that

$$\sum_{k=0}^{m} (k-m) \cdot 10^{k-1} = -\frac{10}{81} 10^m - \frac{m}{9} + \frac{10}{81}.$$

Thus,

$$\sum_{k=0}^{m} a_k \cdot k \cdot 10^{k-1} = \frac{1}{10}nm + O(10^m).$$

But $m = \log n + O(1)$, where $\log n$ denotes the base 10 logarithm of n. Therefore,

$$\sum_{k=0}^{m} a_k \cdot k \cdot 10^{k-1} = \frac{1}{10} n \log n + O(n).$$

Finally,

$$\sum_{k=0}^{m} [a_k > 1] \cdot 10^k + \sum_{k=0}^{m} [a_k = 1] \cdot n_{k-1} = O(n).$$

Therefore, by use of Lemma 2, we have the following corollary.

<u>Corollary</u>.

$$D_1(n) = \frac{1}{10} n \log n + O(n).$$

However, our goal is to examine and more explicitly state the $O(n)$ term.

The Remainder Term. The next step in analyzing $D_1(n)$ is to study it from a different perspective. Let

$$R_1(n) = \frac{1}{10}mn - \sum_{k=0}^{m} a_k \cdot k \cdot 10^{k-1} - \sum_{k=0}^{m-1} [a_k > 1] \cdot 10^k - \sum_{k=0}^{m-1} [a_k = 1] \cdot n_{k-1}.$$

The next lemma will state some properties of R_1, which will be extremely useful throughout the rest of the paper.

Lemma 3.

(a) For any positive integer n, $R_1(10n) = 10 \cdot R_1(n)$.

(b) Let n be a positive integer with decimal representation

$$n = \sum_{k=0}^{m} a_k \cdot 10^k.$$

Then

$$R_1(n+1) - R_1(n) = \frac{m}{10} + [a_m = 1] - d_1(n),$$

where $d_1(n)$ denotes the number of 1's in the decimal representation of n.

(c) Let m be a non-negative integer, d a digit, and p an integer such that $0 \le p < 9 \cdot 10^m$. Then

$$R_1(10^{m+1} + 10p + d) - d \cdot R_1(10^m + p + 1) - (10 - d) \cdot R_1(10^m + p) = \frac{d}{10} - [d > 1].$$

Proof. The proof of (a) involves a fairly easy, but tedious, derivation using the definition of $R_1(n)$. In passing, we note that using Lemma 3(a) and the fact that

$$D_1(n) = \frac{1}{10} mn + [a_m = 1] \cdot n_{m-1} + [a_m > 1] \cdot 10^m - R_1(n),$$

it is immediate that

$$D_1(10n) = 10 \cdot D_1(n) + n \quad \text{for all } n \ge 1.$$

The proof of (b) follows from scrutinizing three cases. The first case is when n and $n+1$ have a different number of digits. Thus $n = 10^{m+1} - 1$. The second case is when n and $n+1$ have the same number of digits but have a different first digit. Thus $n = d \cdot 10^m - 1$ for $d = 2, 3, \ldots, 9$. The third case is the rest of the story, i.e., when n and $n+1$ have the same number of digits and the same first digit. In every case we have that

$$R_1(n+1) - R_1(n) = \frac{m}{10} + [a_m = 1] - d_1(n).$$

The proof of (c) is a little more involved. Using Lemma 3(b) twice, Lemma 3(a) once, and the assumption that the decimal representation of $10^m + p$ is

$$10^m + p = \sum_{k=0}^{m} a_k \cdot 10^k,$$

we have the following sequence of equalities.

$$\begin{aligned}
&R_1(10^{m+1}+10p+d)-d\cdot R_1(10^m+p+1)-(10-d)\cdot R_1(10^m+p)\\
&=R_1(10^{m+1}+10p+d)-d\cdot(R_1(10^m+p+1)-R_1(10^m+p))\\
&\quad -10\cdot R_1(10^m+p)\\
&=R_1(10^{m+1}+10p+d)-R_1(10^{m+1}+10p)\\
&\quad -d\cdot\left(\frac{m}{10}+[a_m=1]-d_1(10^m+p)\right)\\
&=\sum_{k=0}^{d-1}\left(R_1(10^{m+1}+10p+k+1)-R_1(10^{m+1}+10p+k)\right)\\
&\quad -d\cdot\frac{m}{10}-[a_m=1]\cdot d+d\cdot d_1(10^m+p)\\
&=\sum_{k=0}^{d-1}\left(\frac{m+1}{10}+[a_m=1]-d_1(10^{m+1}+10p+k)\right)\\
&\quad -d\cdot\frac{m}{10}-[a_m=1]\cdot d+d\cdot d_1(10^m+p)\\
&=d\cdot\frac{m}{10}+\frac{d}{10}+[a_m=1]\cdot d-\sum_{k=0}^{d-1}d_1(10^{m+1}+10p+k)\\
&\quad -d\cdot\frac{m}{10}-[a_m=1]\cdot d+d\cdot d_1(10^m+p)\\
&=\frac{d}{10}-[d>1].
\end{aligned}$$

This completes the proof of (c) and the proof of Lemma 3.

Next, let m be a non-negative integer and p be an integer such that $0 \le p < 9\cdot 10^m$. Define the function $\phi(x)$ by

$$\phi\left(\frac{p}{9\cdot 10^m}\right)=\frac{R_1(10^m+p)}{10^m}.$$

Note that by Lemma 3(a), $\phi(x)$ is uniquely defined. For if x has any other representation, i.e.

$$\frac{p'}{9\cdot 10^{m'}},$$

then

$$p'=10^{m'-m}p.$$

If we assume, without loss of generality, that $m' > m$, then $m' - m$ is a positive integer. Therefore,

$$\frac{R_1(10^{m'} + p')}{10^{m'}} = \frac{R_1(10^m + p)}{10^m}.$$

The function $\phi(x)$ is defined only on a subset of $[0, 1]$. We now consider the problem of extending this function continuously to $[0, 1]$. Here, we solve this problem by considering the limit of a sequence of "polygonal" functions which identify with $\phi(x)$ on the rationals of the form

$$\frac{p}{9 \cdot 10^m}.$$

These polygonal functions are defined in the following way. Let m be a non-negative integer. $\{f_m(x)\}$ is defined on $[0, 1]$ to be the function whose graph is the polygon joining the points

$$\{(0,0), \left(\frac{1}{9 \cdot 10^m}, \phi(\frac{1}{9 \cdot 10^m})\right), \ldots, \left(\frac{p}{9 \cdot 10^m}, \phi(\frac{p}{9 \cdot 10^m})\right), \ldots, (1,0)\}.$$

Then the definition of $\{f_m(x)\}$ is extended to the reals by $f_m(x \pm 1) = f_m(x)$. From the definition, $f_0(x) \equiv 0$. In addition, $f_1(x)$ is equal to the auxiliary function $g(x)$, which is defined on the reals by

$$g(x) = \begin{cases} \frac{9}{10}x, & 0 \le x < \frac{1}{90}; \\ -\frac{81}{10}x + \frac{1}{10}, & \frac{1}{90} \le x < \frac{2}{90}; \\ \frac{9}{10}x - \frac{1}{10}, & \frac{2}{90} \le x < \frac{1}{9}. \end{cases}$$

and $g(x \pm \frac{1}{9}) = g(x)$. Furthermore, using Lemma 3(c), it follows that for any non-negative integer m and all real x,

$$f_{m+1}(x) - f_m(x) = \frac{1}{10^m} g(10^m x).$$

Repeated iterations of this equation yields

$$f_{m+1}(x) = \sum_{i=0}^{m} \frac{1}{10^i} g(10^i x).$$

Since $g(x)$ is bounded, the sequence $\{f_m(x)\}$ converges uniformly for all x. Hence the limiting function

$$f(x) = \sum_{i=0}^{\infty} \frac{1}{10^i} g(10^i x)$$

is a continuous extension of $\phi(x)$.

The Main Result. Let m be a non-negative integer and p be an integer such that $0 \le p < 9 \cdot 10^m$. In addition, let $n = 10^m + p$ and suppose the decimal representation of n is

$$n = \sum_{k=0}^{m} a_k \cdot 10^k.$$

Then

$$\begin{aligned} D_1(n) &= \frac{1}{10}mn + [a_m = 1] \cdot n_{m-1} + [a_m > 1] \cdot 10^m - R_1(n) \\ &= \frac{1}{10}mn + [a_m = 1] \cdot n_{m-1} + [a_m > 1] \cdot 10^m - 10^m \cdot f(\frac{p}{9 \cdot 10^m}). \end{aligned}$$

Next, since

$$n = 10^m + p = 10^m(1 + \frac{p}{10^m}),$$
$$m = \log n - \log\left(1 + \frac{p}{10^m}\right).$$

Also,

$$n_{m-1} = n - a_m \cdot 10^m = 10^m + p - a_m \cdot 10^m = 10^m\left(1 + \frac{p}{10^m} - a_m\right).$$

Substituting for these two quantities and setting

$$x = \frac{p}{9 \cdot 10^m},$$

we have

$$\begin{aligned} D_1(n) &= \frac{1}{10}n \cdot (\log n - \log(1 + 9x)) \\ &\quad + [a_m = 1] \cdot 10^m(1 - a_m + 9x) + [a_m > 1] \cdot 10^m - 10^m f(x) \\ &= \frac{1}{10}n \log n - \frac{1}{10}n \log(1 + 9x) \\ &\quad + [a_m = 1] \cdot 10^m(1 - a_m + 9x) + [a_m > 1] \cdot 10^m - 10^m f(x). \end{aligned}$$

Finally, substituting $n = 10^m(1 + 9x)$ for the second n, we restate the result in the following theorem.

Theorem. Let n be a positive integer with decimal representation

$$n = \sum_{k=0}^{m} a_k \cdot 10^k.$$

In addition, let $n = 10^m + p$ and

$$x = \frac{p}{9 \cdot 10^m}.$$

Then

$$D_1(n) = \frac{1}{10} n \log n - 10^{m-1} \biggl(10 f(x) + (1 + 9x) \log(1 + 9x) - [a_m > 1] \cdot 10 - [a_m = 1] \cdot 90x \biggr).$$

Questions. We end this paper with some questions for further work. First of all, we would like to find an expression for the number of 0's, or 2's, or 3's, etc. in the decimal representation of the sequence of positive integers less than n. We feel that this question is assessible using the techniques in this paper. Another question is the matter of different bases. That is, can we find a formula for the number of 1's in the base b representation of the sequence of positive integers less than n. Therefore, in general, we would like to determine a formula for the number of digits d in the base b representation of the sequence of positive integers less than n.

References

1. J. Trollope, "Generalized Bases and Digital Sums," *American Mathematical Monthly,* 74 (1967), 690–694.
2. J. Trollope, "An Explicit Expression for Binary Digital Sums," *Mathematics Magazine,* (1968), 21–25.
3. R. Graham, D. Knuth, and O. Patashnik, *Concrete Mathematics,* Addison-Wesley Publishing Company, Reading, MA, 1989.