Distributions of Letter Usage in the Names of Positive Integers

Matthew Schaffer

23 February 2019

Introduction. We seek to understand if there is a limit in the distribution of letter usage in the names of the positive integers. We begin by determining the formula for counting average letter usage in the names of the positive integers from 1 to N, including 1 but not N. To simplify the task, we will calculate the average letter usage for the names of the positive integers up to, but not including, $10^{(3K+3)}$, for $K \geq 0$. It is the limit of the frequencies determined by these average distributions that we seek.

Along the way, we introduce labeling conventions for very large powers of ten—more than one convention exists, and the convention choice matters. We then calculate a formula for the average letter usage for a class of very large powers of ten, and we determine this limit. We then discuss whether the limit thus calculated is, in fact, the limiting distribution of the letter frequency in the names of the positive integers up to $10^{(3K+3)}$.

Definitions. We start with the U.S. short-scale[1] labeling[2] convention for integer powers of ten, as shown in Table 1.

Table 1. Short-scale labeling convention for powers of 10 up to 10^{33}

Value in scientific notation	Label
10^0	One
10^1	Ten
10^2	Hundred
10^3	Thousand
10^4	Ten thousand
10^5	Hundred thousand
10^6	Million
10^7	Ten million
10^8	Hundred million
10^9	Billion
10^{12}	Trillion
10^{15}	Quadrillion
10^{18}	Quintillion
10^{21}	Sextillion
10^{24}	Septillion
10^{27}	Octillion
10^{30}	Nonillion
10^{33}	Decillion

[1] See, for example, https://en.wikipedia.org/wiki/Long_and_short_scales#Short_scale and https://en.wikipedia.org/wiki/Names_of_large_numbers#cite_note-a-14

[2] We differentiate between *names* of numbers and *labels* of powers of ten as follows: 1,000 is *named* "one thousand" and its *label* is "thousand."

We use the standard English names for the numbers from 1 to 100, e.g., one, five, eleven, fifteen, twenty one, twenty five, thirty one, etc. The total letter counts for the positive integers up to (but not including) 10 is shown in Figure 1, and up to (but not including) 100, in Figure 2. We generate these totals essentially by brute force—writing out the names and adding up the individual letters. A spreadsheet simplifies the task.

Figure 1. Total letter counts for positive integers up to 10

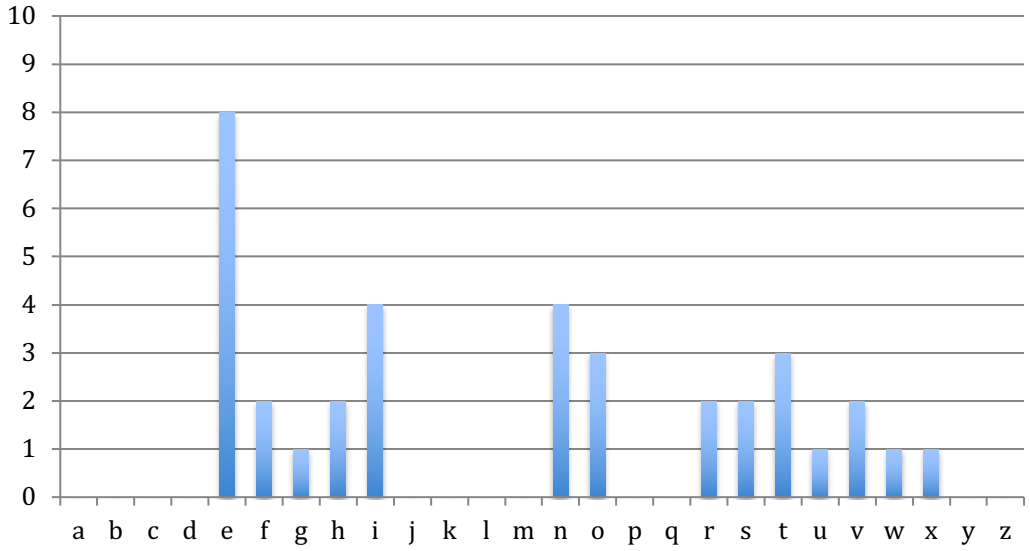
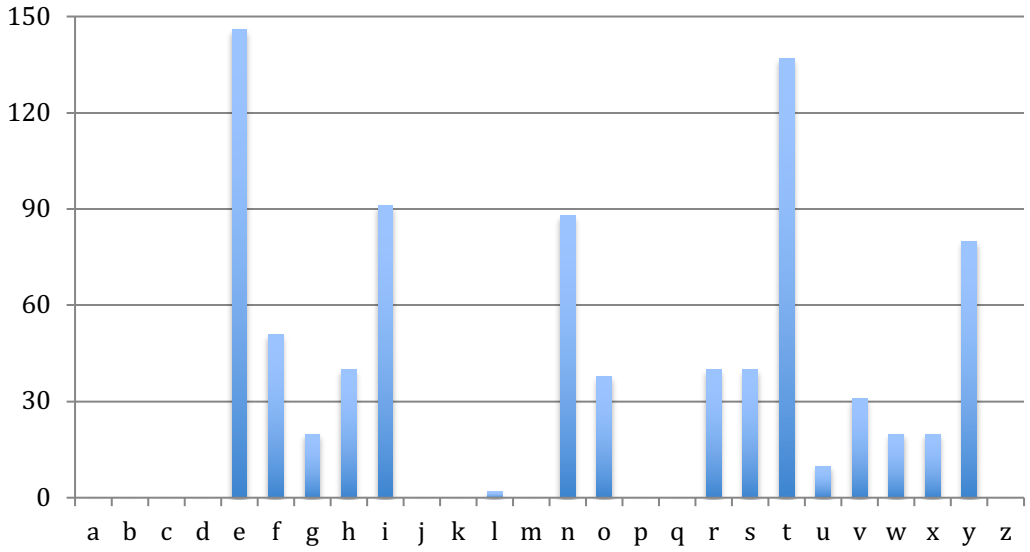


Figure 2. Total letter counts for positive integers up to 100



Underlying formula. As we head to one thousand, we begin to gain insight into the formula we will rely on extensively. Almost every number between 100 and 999, inclusive, can be named as shown in Table 2.

Table 2. Rule for names of integers between 100 and 999, inclusive

First term	Second term	Third term
Choose a name of a number from the set {1, 2, … 9}	Add the label "hundred"	Add a name of a number from the set {1, 2, … 99}

The exceptions to this rule are, of course, the exact multiples of one hundred, which have no third term. They are named, simply, one hundred, two hundred, etc.

Now, define $N(\lambda,i)$ to be the number of occurrences of the letter $\lambda \in \{a, b, c, \ldots z\}$ in the names of the positive integers up to, but not including, 10^i, for $i \geq 1$. Similarly, let $L(\lambda,j)$ be the number of occurrences of the letter λ in the label for 10^j, for $j \geq 2$, and let $A(\lambda,i)$ be the average number of occurrences of the letter λ in the names of the positive integers up to, but not including 10^i. Specifically[3], $A(\lambda,i) = N(\lambda,i)/10^i$.

Using these definitions and Table 2, we find

(Eq. 1) $$N(\lambda,3) = 10 * N(\lambda,2) + 100 * N(\lambda,1) + 9 * 100 * L(\lambda,2)$$

In other words, to find the total number of λs found in the names of the positive integers up to 1,000, we repeat ten times all the λs in the integers between 1 and 99 (inclusive)—one for each multiple of 100, to include the zero multiple—which accounts for all the third terms in Table 2. We repeat 100 times all the λs in the integers between 1 and 9 (inclusive), which accounts for all the first terms in the table. Finally, we repeat 900 times the label "hundred," which accounts for all the second terms.

We now start jumping three orders of magnitude in our calculations, and we'll develop a recursive formula. As we head to one million, we rely on the naming rule for every number between 1,000 and 999,999 shown in Table 3 (again, excepting that the exact multiples of 1,000 have no third term).

[3] OK, the witting reader will realize this is not quite right. The denominator should really be 10^i-1. But we're headed to infinity, so close enough. Alternatively, include the letters in the name "zero" in the definition of the $N(\lambda,i)$, and now the equation is exactly right. It will make no difference in the limiting case.

Table 3. Rule for naming integers between 1,000 and 999,999, inclusive

First term	Second term	Third term
Choose a name of a number from the set {1, 2, … 999}	Add the label "thousand"	Add a name of a number from the set {1, 2, … 999}

As above, we find:

(Eq. 2) $N(\lambda,6) = 10^3 * N(\lambda,3) + 10^3 * N(\lambda,3) + 999 * 10^3 * L(\lambda,3)$

In other words, to find the total number of λs found in the names of the positive integers up to, but not including, 1,000,000, we repeat one thousand times all the λs in the names of the integers between 1 and 999 (inclusive)—one for each multiple of 1,000, to include the zero multiple—which accounts for all the third terms. We repeat 1,000 times all the λs in the names of the integers between 1 and 999 (inclusive), which accounts for all the first terms. Finally, we repeat 999,000 times the label "thousand," which accounts for all the second terms.

We find the naming rule for positive integers up to, but not including, 10^{3K} in Table 4, and the recursive formulae for the corresponding total and average letter counts at Equations 3 and 4, respectively. The equations arise using the same logic as we used in moving from Table 3 to Equation 2. Specifically, we repeat all the third terms one thousand times; we repeat the first terms $10^{3(K-1)}$ times; and we repeat $999 * 10^{3(K-1)}$ times the label for $10^{3(K-1)}$.

Table 4. Rule for naming integers between $10^{3(K-1)}$ and $10^{3K} - 1$, inclusive

First term	Second term	Third term
Choose a name of a number from the set {1, 2, … 999}	Add the label corresponding to $10^{3(K-1)}$	Add a name of a number from the set {1, 2, … $10^{3(K-1)} - 1$}

(Eq. 3) $N(\lambda,3K) = 10^3 * N(\lambda,3(K-1)) + 10^{3(K-1)} * N(\lambda,3) + 999 * 10^{3(K-1)} * L(\lambda,3(K-1))$

(Eq. 4) $A(\lambda,3K) = A(\lambda,3(K-1)) + A(\lambda,3) + 0.999 * L(\lambda,3(K-1))$

Using Equation 4 recursively, we find a simple formula for the average letter count in the names of positive integers, namely Equation 5

(Eq. 5) $A(\lambda,3K) = A(\lambda,3) + (K-1) * A(\lambda,3) + 0.999 * \sum_{j=1}^{K-1} L(\lambda, 3j)$

Equation 5 is the underlying formula for our task. We define the frequency of occurrence of the letter λ in the names for positive integers up to, but not including, 10^{3K}, as F(λ,K). We find this frequency simply by dividing A(λ,K) with the sum of, over all λs from a to z, the terms A(λ,K). Our goal, then, is to find the frequency limit

$$\lim_{K\to\infty} F(\lambda, K) = \lim_{K\to\infty} \frac{A(\lambda, 3K)}{\sum_\lambda A(\lambda, 3K)}$$

If we divide Equation 5 by (K-1) and ignore the first summand on the right side of the equation, we find that the frequency limit reduces to

$$\lim_{K\to\infty} F(\lambda, K) = \lim_{K\to\infty} \frac{A(\lambda, 3) + 0.999 * \text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1)}{\sum_\lambda [A(\lambda, 3) + 0.999 * \text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1)]}$$

As we will see, the label averages for most λs[4] are going to grow without bound, which means we need only concern ourselves with the right-hand terms in our limit—in both the numerator and denominator. Thus, our task quickly reduces to understanding the behavior of the cumulative averages of the number of λs in the labels for powers of 10, or, more simply, we seek the following limit

$$\lim_{K\to\infty} F(\lambda, K) = \lim_{K\to\infty} \frac{\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1)}{\sum_\lambda \text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1)}$$

Labeling conventions. For numbers less than 10^{33}, the labeling convention for the second term in Table 4 is found in Table 1. Above 10^{33}, we introduce the next of our labeling conventions. The label of the number $10^{(3K+3)}$, for K = 10, 11, … 99, is obtained by concatenating Latin roots (modified as explained in the notes) for the

[4] Three letters (j, k, and z) never appear. Four others (f, h, w, and y) do not appear in the labels of equation 5. For these latter cases, the numerator becomes fixed—specifically, A(λ, 3)—but the denominator grows unbounded. Thus, these seven frequency limits are all 0.

units and tens place of K, in that order, followed by the suffix "llion." Table 5[5] contains the Latin roots. (Note that, when the unit place is a zero, no unit root is included in the label.)

Table 5. Short-scale Latin roots for labels of powers of 10 between 33 and 300, inclusive

	Units	Tens
0		
1	un	deci [N]
2	duo	viginti [MS]
3	tre[6]	triginti [NS]
4	quattuor	quadraginti [NS]
5	quin	quinquaginti [NS]
6	se[7]	sexaginti [N]
7	septe[8]	septuaginti [N]
8	octo	octoginti [MX]
9	nove[9]	nonaginti

This system was developed by John Horton Conway and Allan Wechsler, and then refined by Olivier Miakinen. Table 6 illustrates this labeling convention for a few examples of powers of ten.

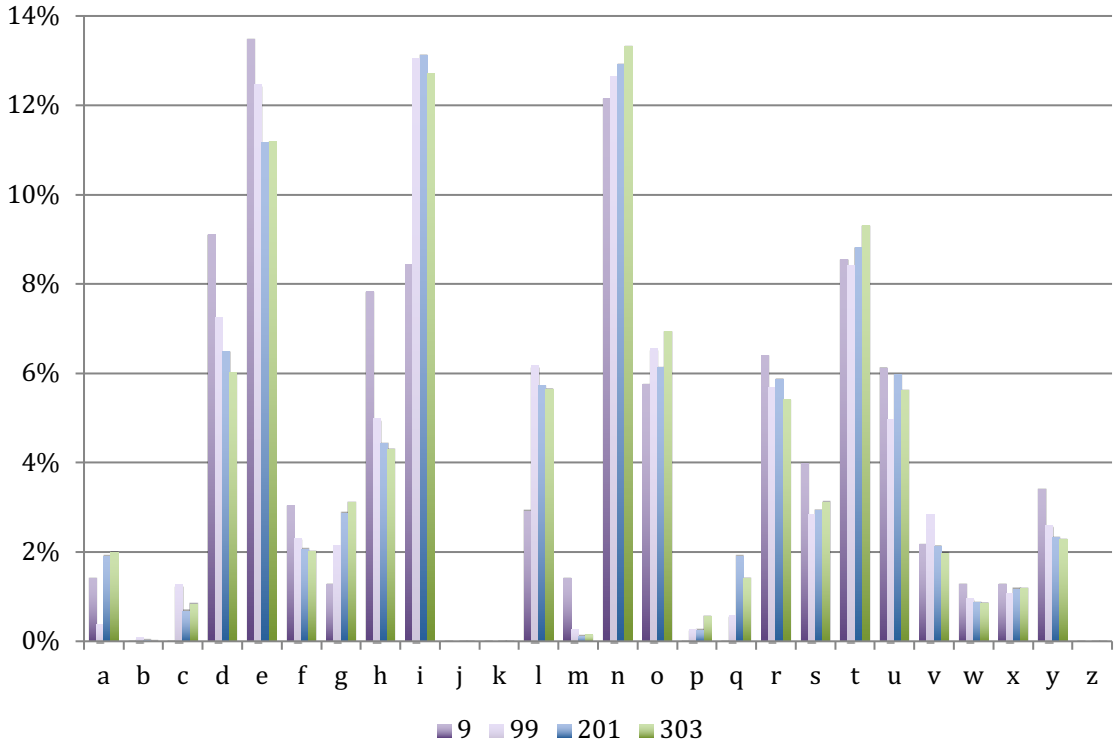
Table 6. Example labels of powers of 10 between 33 and 303

10^{3N+3}	N	Label
10^{36}	11	undecillion
10^{39}	12	duodecillion
10^{42}	13	tredecillion
10^{63}	20	vigintillion
10^{72}	23	tresvigintillion
10^{84}	27	septemvigintillion
10^{111}	36	sestrigintillion
10^{120}	39	noventrigintillion
10^{180}	59	novenquinquagintillion
10^{261}	86	sexoctogintillion
10^{300}	99	novenonagintillion

[5] See, for example, http://mrob.com/pub/math/largenum.html#chuquet
[6] If the tens component of K is marked with an S or X, then "tre" changes to "tres"
[7] If the tens component of K is marked with an S or X, then "se" changes to "ses" or "sex"
[8] If the tens component of K is marked with an M or N, then "septe" changes to "septem" or "septen"
[9] If the tens component of K is marked with an M or N, then "nove" changes to "novem" or "noven"

While we show only labels for the powers of ten divisible by three, we note the intermediate powers of 10 use the naming convention specified in Table 4. Thus, 10^{36} is named one undecillion, 10^{37} is ten undecillion, and 10^{38} is one hundred undecillion. Figure 3 shows the first of our letter-frequency distributions for large powers of 10.

Figure 3. Percent letter frequency in names of the positive integers Between 1 and 10^N - 1, inclusive, for N = 9, 99, 201, and 303



Compare this figure to Figure 2. In that figure, "e" and "t" were the most frequent letters. Now, "i" and "n," which were the third and fourth most frequent, are the two most frequent letters for the higher powers of 10. Note also the significant increase in the frequency of the letters "l" and "o." These increases all result from the "llion" suffix being so prevalent in the labels of higher powers of 10. Notice also that the frequencies for "f," "h," "w," and "y" are still sizeable but decaying—they are all headed to zero as noted earlier. And, while "b" appears to have a zero frequency, it does not, though it is very small and will stay small.

The next step in our labeling convention will take us all the way to $10^{3,003}$. It is the second part of the Conway-Wechsler convention. Now, the Ks in 10^{3K+3} include a "hundreds" place. As before, the label of the number $10^{(3K+3)}$, for K = 100, 101, … 999, is obtained by concatenating Latin roots for the units, tens, and hundreds place of K, in that order, followed by the suffix "llion." Note that most of the tens roots

now end in "a" rather than "i." Table 7[10] contains the Latin roots. Note also that, should either the unit or ten (or both) place equal zero, no root is included for that place.

Table 7. Short-scale Latin roots for labels of powers of 10 between 303 and 3,000, inclusive

	Units	Tens	Hundreds
0			
1	un	deci [N]	centi [NX]
2	duo	viginti [MS]	ducenti [N]
3	tre[11]	triginta [NS]	trecenti [NS]
4	quattuor	quadraginta [NS]	quadringenti [NS]
5	quin	quinquaginta [NS]	quingenti [NS]
6	se	sexaginta [N]	sescenti [N]
7	septe	septuaginta [N]	septingenti [N]
8	octo	octoginta [MX]	octigenti [MX]
9	nove	nonaginta	nongenti

We offer only one example using Table 7. The name for $10^{1,365} = 10^{3*454+3}$ is one quattuorquinquagintaquadringentillion (quattuor-quinquaginta-quadringenti-llion). The name is interesting only as it is longest label for a power of 10 thus far, which is easy to see by inspection of Table 7 (longest unit, ten, and hundred roots).

Now that we have reached $10^{3,003}$, we introduce a new, and final, labeling convention developed by Alan Wechsler[12], still jumping three orders of magnitude. We express the power of 10 as 10^{3N+3}. We further express N, for $N \geq 1{,}000$, as

(Eq. 6) $$N = X_0 + X_1 * 10^3 + X_2 * 10^6 + \ldots + X_M * 10^{3M}, \text{ for } M \geq 1.$$

Each X_j (j = 0, 1, 2, … M) is a number between 0 and 999, inclusive, except for X_M, which must be greater than 0. The labeling convention uses the prefixes (i.e., the letters prior to "llion") developed thus far for the labels of 10^{3X_j+3}, for all j. To each of these prefixes, we concatenate "lli." Should $X_j = 0$, for any $0 \leq j < M$, we use "ni" as

[10] See, for example, http://mrob.com/pub/math/largenum.html#chuquet

[11] If the adjoining (either tens or hundreds) component of K is marked with an S or X, then the units component "tre" changes to "tres." Use only the rule for the adjoining component. For example, for N=103, roots are tres-centi. For N=113, roots are tre-deci-centi. Likewise for units 6, 7, and 9, and following the same footnotes for Table 5. For example, for N=106, roots are sex-centi. For N=116, roots are se-deci-centi.

[12] See, for example, http://mrob.com/pub/math/largenum.html#cw_beyond_lp1_c3000

the prefix. For all other X_j, we use Table 1 to find the prefixes for 10^{3X_j+3} for X_j = 1 through 10, then Tables 5 and 7 for X_j = 11 through 999.

Now, define Prefix(X_j) to be the prefix corresponding to 10^{3X_j+3}, for X_j = 1, 2, … 999. Define Prefix(0) = "ni." To construct the label for 10^{3N+3}, we start with the prefix for X_M, add "lli," then the prefix for X_{M-1}, add "lli", and so on, ending with the prefix for X_0 and the final "lli." We then concatenate "on" at the very end.

For example, Prefix(1) = "mi," which is the prefix corresponding to the label for 10^{3*1+3}, or million. Likewise, Prefix(39) = "noventriginti," obtained from Table 5 (or Table 6). Finally, Prefix(454) = "quattuorquinquagintaquadringenti."

With the above definitions, the label for 10^{3N+3}, using Equation 6, would be

(Eq. 7) Prefix(X_M) + "lli" + Prefix(X_{M-1}) + "lli" + … + Prefix(X_0) + "lli" + "on",

where the "+"s in Equation 7 mean concatenate.

So, for X_1 = 1 and X_0 = 0, we get mi-lli-ni-lli-on, or millinillion, which is the label for $10^{3*1,000+3}$ or $10^{3,003}$. For X_1 = 1 and X_0 = 1, we get mi-lli-mi-lli-on, or millimillion as the label for $10^{3,006}$. For X_2 = 6, X_1 = 454, and X_0 = 17, we get sexti-lli-quattuorquinquagintaquadringenti-lli-septendeci-lli-on or sextilliquattuorquinquagintaquadringentilliseptendecillion, which is the label for $10^{3*(6*1,000,000+454*1,000+17)+3}$ or $10^{19,362,054}$.

Figure 4 shows a few distributions of these higher powers of 10. We illustrate with these frequency distributions the results using the labeling conventions in Tables 5 and 7 for the powers up to $10^{3,003}$ and only the very beginnings of the results using the labeling conventions defined by Equation 7. Note still a sizable variance in the distribution results over this modest range of powers of 10.

In addition, we show, in Figure 5, the sum, over all λs, of the cumulative averages of letters in the labels embedded in the results in Figure 4. In other words, Figure 5 plots the right-hand term in the denominator of our frequency limit shown earlier. Recall that understanding the behavior of the cumulative averages expressed in Equation 5 is critical to understanding whether the frequency limit exists.

Note that Figure 5 suggests some stability and, therefore, lays hope for convergence of our frequency limit. But we have many powers of 10 to go.

Figure 4. Percent letter frequency in names of positive integers between 1 and 10^N - 1, inclusive, for N = 303, 1203, 2103, 3003, 3903, and 6003

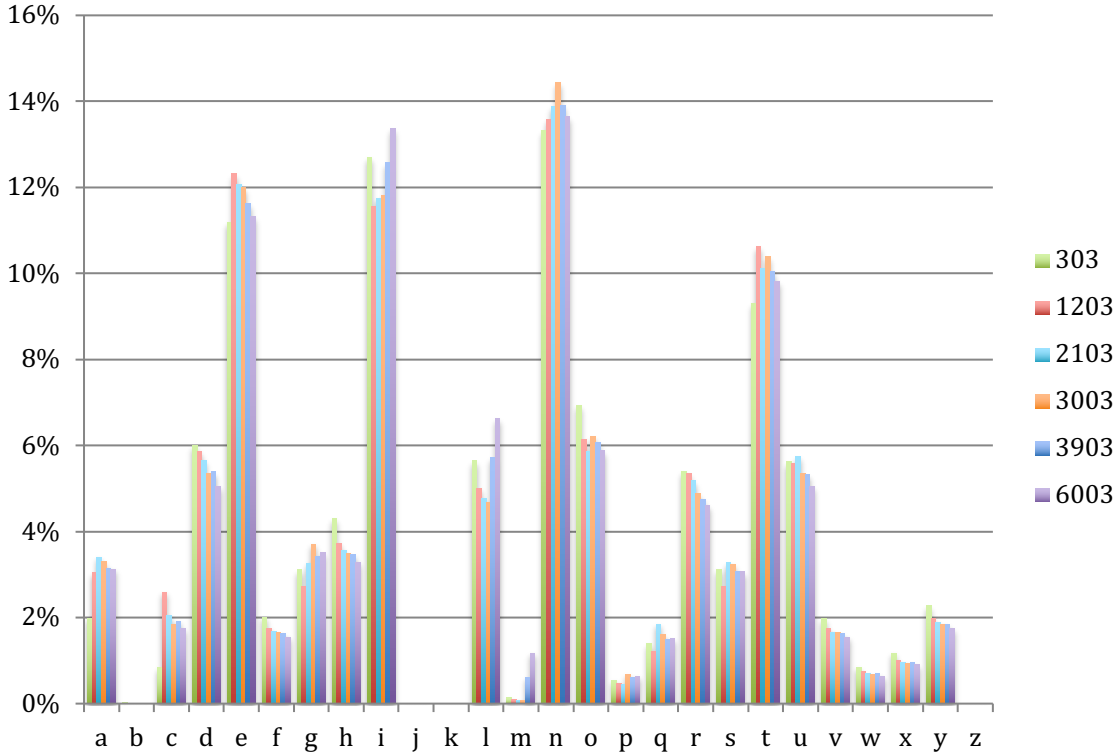
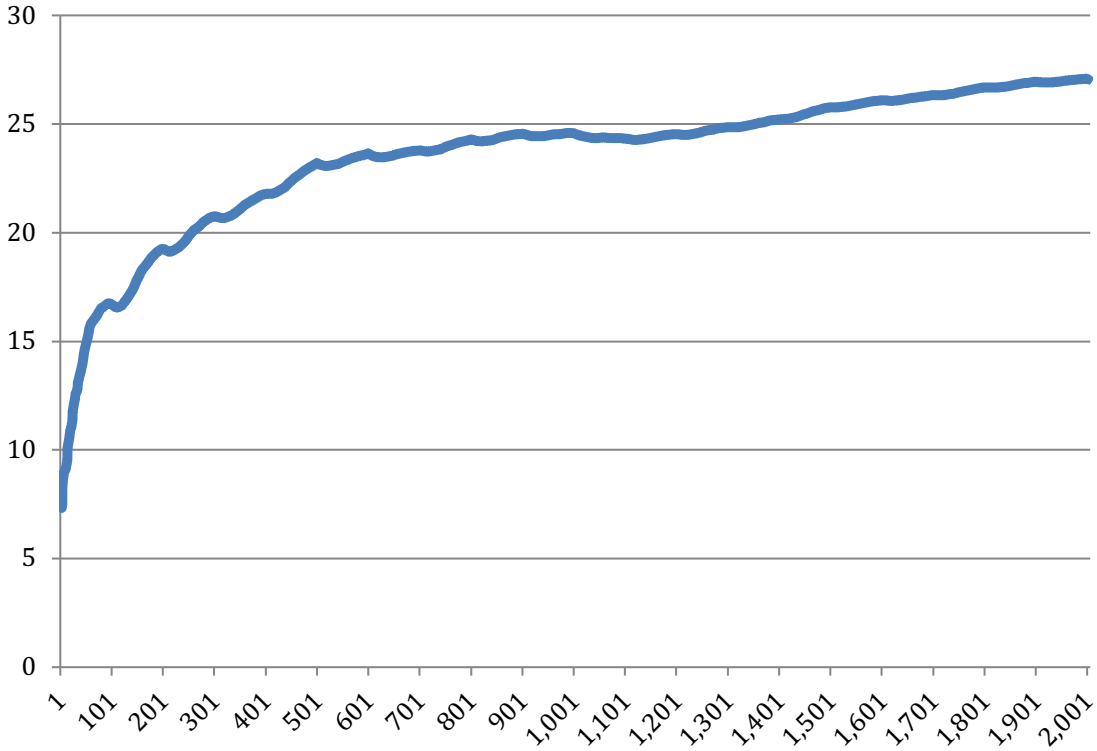


Figure 5. Sum of cumulative average number of letters in labels of powers of ten of the form 10^{3N+3}, for N up to 2,001



Moving toward a limit. The second-to-last step in our effort is to determine a closed-form solution for the cumulative averages of labels of powers of ten of the form 10^{3K+3} per Equation 5. That is, we want to understand how

(Eq. 8) $$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1) = \frac{\sum_{j=1}^{K-1} L(\lambda,3j)}{K-1}$$

behaves as K tends to ∞.

We start by looking at the cases where $K = 10^{3M}$, for $M \geq 2$. Note the size of the jumps we are now taking. We set out to find the limit of letter frequency as we leapt by three orders of magnitude, i.e., 10^3, 10^6, 10^9, etc. Now, we are leaping by three orders of magnitude in the exponent.

First, let $\mathcal{L}(\lambda) = \sum_{i=1}^{1,000} L(\lambda,3i)$, which is the total count of the letter λ in labels of the integers between 10^3 and $10^{3,000}$, inclusive. Then Equation 8 becomes, for $K \geq 1{,}002$

(Eq. 9) $$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \dots, K-1) = \frac{\mathcal{L}(\lambda)}{K-1} + \frac{\sum_{j=1001}^{K-1} L(\lambda,3j)}{K-1}$$

Let $K = 10^6 + 1$ in Equation 9. To simplify the final summand in Equation 9, we examine Table 8, which shows the labeling convention for the labels between $10^{3,003}$ (millinillion) and $10^{3,000,000}$ (novenonagintanongentillinovenonagintanongentillion) following the convention specified in Equation 6—namely, each label is a concatenation of a "thousands" term, a "units" term.

Table 8. Rule for labeling powers of 10^{3N+3} between $10^{3,003}$ and $10^{3,000,000}$

Choose one "thousands" term, X_1 +"lli"	Choose one "units" term, X_0 + "lli"	End with
	nilli	on
milli	milli	
billi	billi	
trilli	trilli	
…	…	
novenonagintanongentilli	novenonagintanongentilli	

Now, let $P_1(\lambda)$ be the sum of all λs in the "thousands" terms (including the "lli") for the labels between $10^{3,003}$ and $10^{3,000,000}$, inclusive, i.e., "milli", "billi", etc., up to "novenonagintanongentilli," as seen in the first column of Table 8. Let $P_0(\lambda)$ be $P_1(\lambda)$

plus the letters in the term "nilli," which represents the sum of all λs in the "units" terms for the labels between $10^{3,003}$ and $10^{3,000,000}$, inclusive, as seen in the second column of the table. Note that there are 999 "thousands" terms and 1,000 "units" terms. Note also that every combination is used once and all 999,000 combinations end with "on."

It is therefore straightforward to see that

$$\sum_{i=1,001}^{1,000,000} L(\lambda,3i) = 10^3\,[P_1(\lambda) + 0.999 * P_0(\lambda) + 999\ (\text{"o"s and "n"s})]$$

Of course, the last term is included in the summation only when λ equals "o" or "n." Substituting into Equation 9 with $K = 10^6 + 1$, we find that

(Eq. 9a)
$$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \ldots 10^6) = \frac{\mathcal{L}(\lambda)}{10^6} + 10^3 * [P_1(\lambda) + 0.999 * P_0(\lambda) + 999\ (\text{"o"s and "n"s})]/10^6$$

Now, let $K = 10^9 + 1$ in Equation 9. Mimicking what we did to get to Equation 9a, we find that

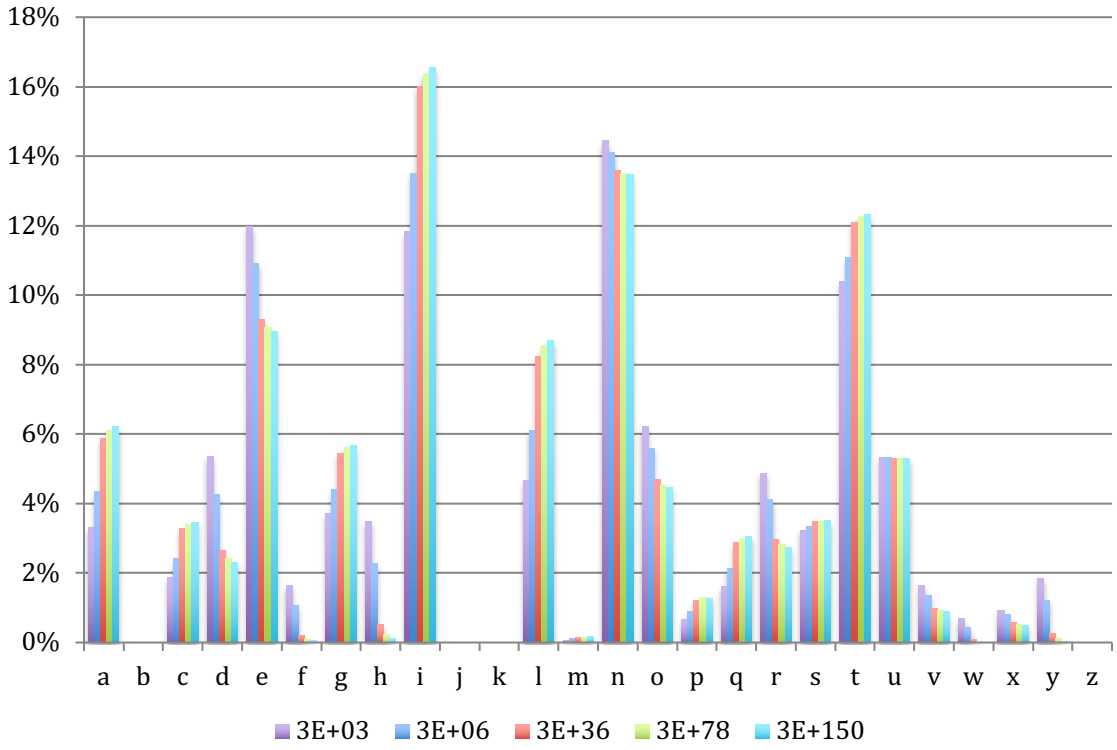
(Eq. 9b)
$$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \ldots 10^9) = \frac{\mathcal{L}(\lambda)}{10^9} + \{10^3 * [P_1(\lambda) + 0.999 * P_0(\lambda) + 999\ (\text{"o"s and "n"s})] + 10^6 * [P_1(\lambda) + 2 * 0.999 * P_0(\lambda) + 999\ (\text{"o"s and "n"s})]\}/10^9$$

Finally, let $K = 10^{3M} + 1$ in Equation 9. Then, with proof left to the reader,

(Eq. 10)
$$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \ldots 10^{3M}) = \frac{\mathcal{L}(\lambda)}{10^{3M}} + [P_1(\lambda) * \sum_{j=0}^{M-2} 10^{3(j+1)} + 0.999 * P_0(\lambda) * \sum_{j=0}^{M-2}(j+1) * 10^{3(j+1)} + 999 * \sum_{j=0}^{M-2} 10^{3(j+1)}\ (\text{"o"s and "n"s})]/10^{3M}$$

Figure 6 provides plots of letter frequencies for such large K.

Figure 6. Percent letter frequency in names of positive integers between 1 and 10^{3K+3} - 1, inclusive, for K = 10^3, 10^6, 10^{36}, 10^{78}, and 10^{150}



Returning to Equation 10 and ignoring the trivial terms, we find, after considerable algebra[13], that

(Eq. 11) $$\text{Average}(L(\lambda, 3j) \mid j = 1, 2, \ldots 10^{3M}) \approx 1/999 * P_1(\lambda) + 1 \text{ (“o”s and “n”s)} + (M - 1)/1000 * P_0(\lambda)$$

As M tends to ∞, only the last term contributes to the frequency of the letter count, which means the limiting frequency distribution is simply the frequency distribution of $P_0(\lambda)$, that is

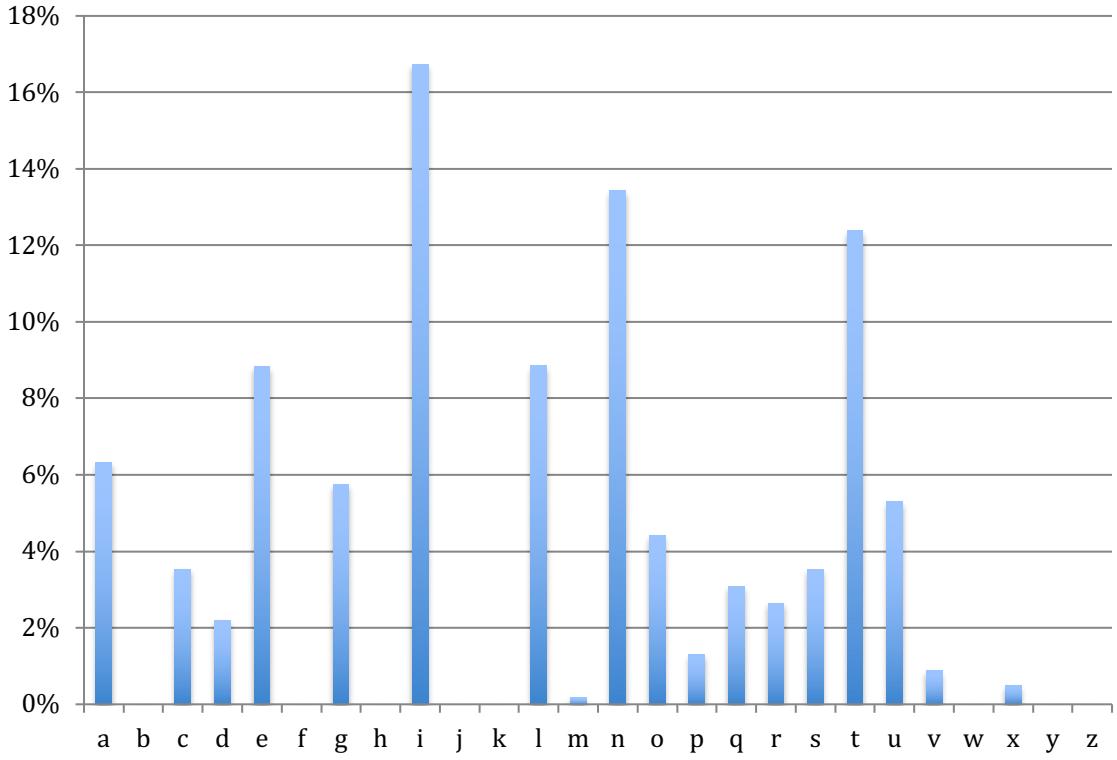
(Eq. 12)

$$\lim_{K \to \infty} F(\lambda, K) = \frac{P_0(\lambda)}{\sum_\lambda P_0(\lambda)}$$

[13] See the appendix for math on reducing Equation 10.

Figure 7 shows the limiting distribution defined by the right side of Equation 12.

Figure 7. Limiting distribution of letter frequency



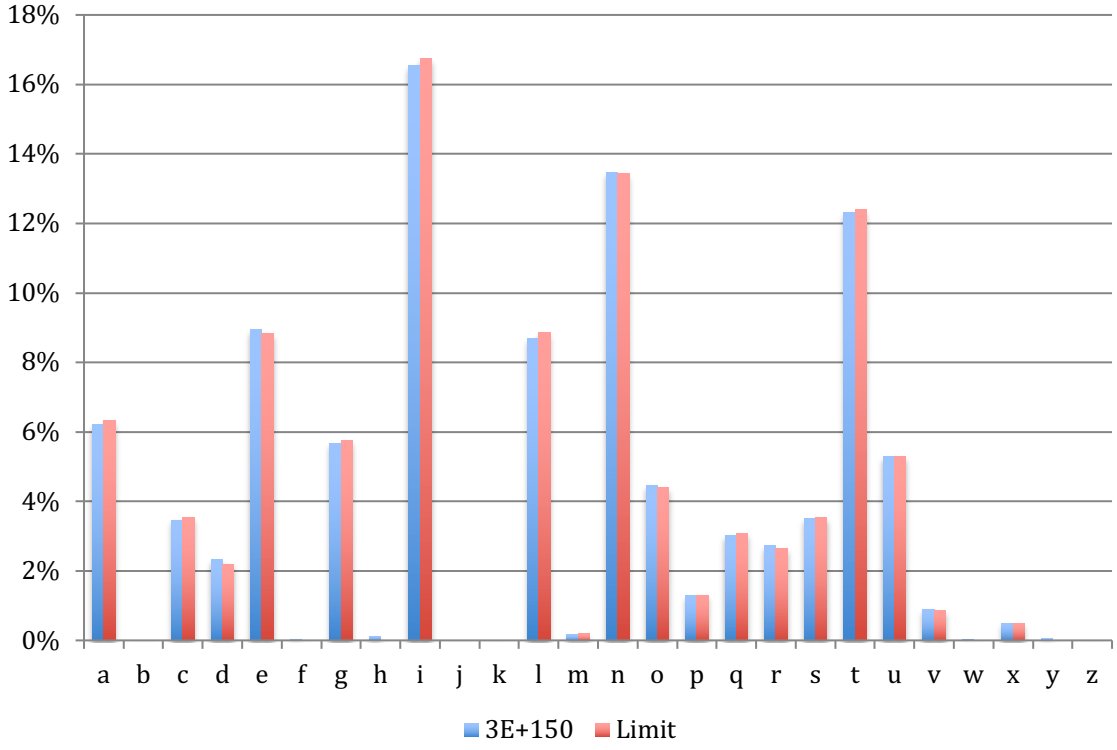
Ensuring a limit. But does the limit in Equation 12 actually exist? We can calculate the right side of the equation, certainly. And, if the limit on the left side of the equation does exist, it must equal the right side.

We made a simplifying assumption to get to this limiting distribution, in effect taking the limit of a subset of the overall sequence of distributions. As we stated in the introduction, we sought to calculate the frequency in the letter composition of the names of the positive integers up to, but not including, $10^{(3K+3)}$, for $K \geq 0$. We allowed jumps of three orders of magnitude when calculating these frequency distributions, expecting some variance in the frequency distributions between the jumps.

Yet, the limiting distribution in Figure 7 arose by jumping three orders of magnitude in the exponents of the powers of ten. That is, we jumped from calculating distributions of letter frequency for positive integers up to 10^{3K+3} to calculating distributions of letter frequency for positive integers up to $10^{3*10^{3M}+3}$. Figure 8 compares the limiting distribution to the distribution computed for $10^{3*10^{3*50}+3}$

using Equation 10. (By the way, that's one millinillinilli…nillion, with 50 consecutive "nilli"s.) Small differences appear for many of the letters, though they are less than 0.2% in all cases. We also note the limiting distribution is either greater or lesser than the computed distribution consistent with the trends suggested by Figure 6.

Figure 8. Comparison of limiting distribution to last calculated distribution



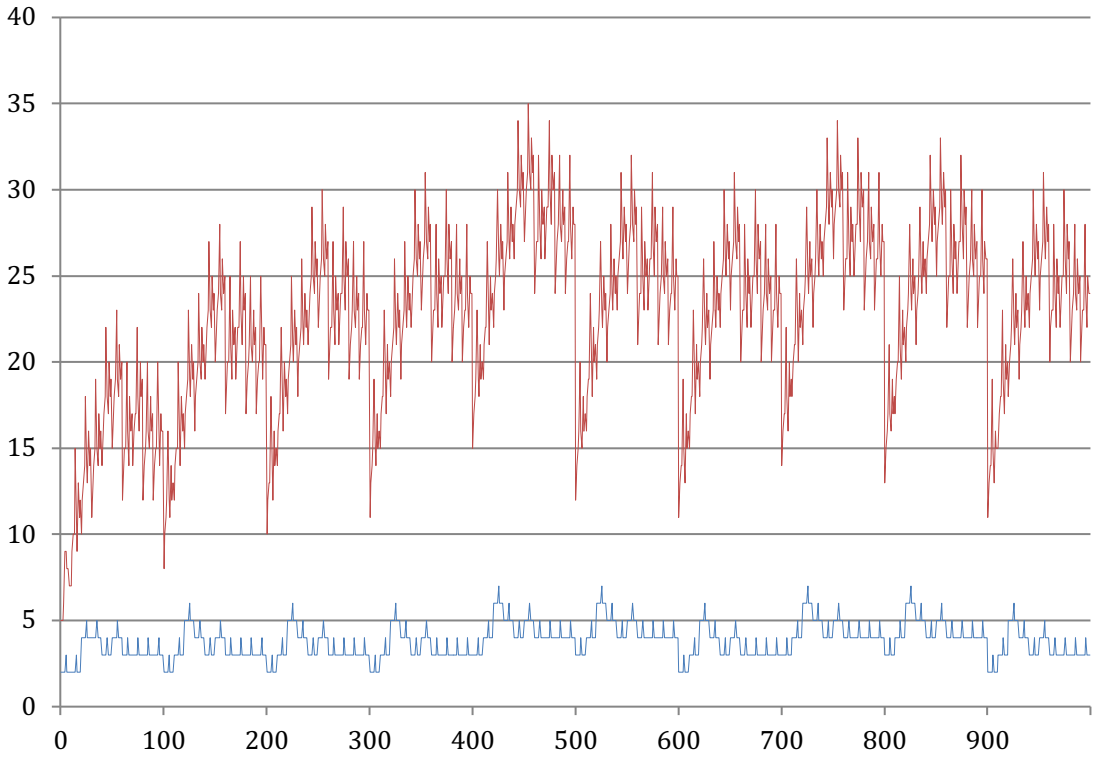
To confirm the existence of a limit, we would need to understand the size of the variance in the intermediary frequency distributions between $10^{3*10^{3(M-1)}+3}$ and $10^{3*10^{3M}+3}$, stepping incrementally by three orders of magnitude. It is conceivable that the subsequence limit exists but the overall limit does not.

We should also note that the cumulative average length of our labels is growing steadily. The approximate average label length, using Equation 11, steadily increases by about 23 letters when we move from $10^{3*10^{3(M-1)}+3}$ to $10^{3*10^{3M}+3}$ for every M. This arises simply because the average, over all λ, of $10^{-3} * P_0(\lambda)$ is 23.

But how much variance is there in between? We see significant growth in our label length, and we know it is not steadily increasing toward this average, as we saw in Figure 5. The prefix names defined by the columns in Table 8 (and its higher-order versions) are the source of the variability in label length.

Figure 9 illustrates this variability. While this figure shows considerable fluctuations in prefix length and the number of "i"s, the question is: how does this variability affect the cumulative average of our label counts for the individual letters?

Figure 9. Variability in prefix length (top) and number of "i"s (bottom) from nilli to novenonagintanongentilli



We choose not to prove with rigor that the limit exists. Instead, we look at the particular case of the frequency of "i"s. By the time we reach $10^{3*10^{3*50}+3}$ (i.e., the case where M = 50 in Equation 11), our simple estimate of the cumulative average of number of "i"s is pretty good—we are within 0.1% of the actual average. The cumulative average number of "i"s is 189, and the sum over all letters of the cumulative averages is 1,132—for an estimate of roughly 17% frequency of the number of "i"s in the positive integer names up to $10^{3*10^{3*50}+3}$.

Our estimate for the number of "i"s in our cumulative average is taken from Equation 11 and is

$$P_1(i)/999 + (50-1)/1000 * P_0(i) = 3778/999 + .049 * 3780 \approx 189$$

Now, as we head toward $10^{3*10^{3*51}+3}$, what happens? In simple terms, we add another column of prefixes, ala Table 8, to the left side of our labels, starting with "milli" and ending with "novenonagintanongentilli." To the right side of these prefixes, we have all the labels for powers of 10 below $10^{3*10^{3*50}+3}$, for which we know the cumulative average number of "i"s.

We know the introduction of the prefix "milli" on the left of our label brings at most 2 more "i"s for each increase of three orders of magnitude in our power of 10, but we don't make a dent in our cumulative average without introducing many of these higher-order labels. We also know we will bring, on average, about 4 more "i"s with all the new prefixes, just by observation of the blue curve at the bottom of Figure 9. And we know we won't introduce more than 7 more "i"s with the new prefixes, again by observation of that same blue curve.

In a sense, we're adding about 4 ± 3 new "i"s to our cumulative averages at each step of three orders of magnitude in our power of 10. While we don't know what the actual effect is on the cumulative average at each step, we do know we'll end up adding about 4 more "i"s by the time we hit $10^{3*10^{3*51}+3}$, and we can estimate the variability by dividing ± 3 by the cumulative average we had already calculated, which was 189. This gives us an estimate of the bound on the variability in our average of about $\pm$ 1.5%. As M tends to ∞, this estimated bound on the variability will head toward 0, which would guarantee convergence of our limit (a similar argument applies to the denominator in Equation 12). Not rigorous but the gist.

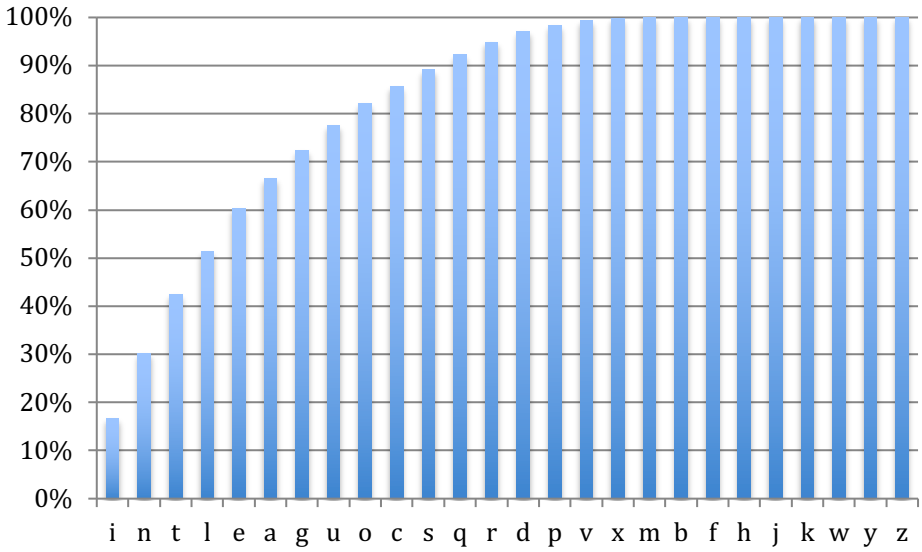
An obvious question is whether the limiting distribution we calculated would be the same had we chosen the long-scale convention (or its alternative) for naming numbers. The answer is a resounding "NO." We would follow the naming convention logic exactly as developed in this paper, but the particulars would differ in a few places.

Specifically, in the long-scale convention, we introduce new labels in steps of a million instead of steps of a thousand, as in the short-scale convention. Thus, 10^9 is named one thousand million (or milliard in the alternative) and 10^{12} is named one billion. One centillion occurs at 10^{600} instead of 10^{303}, as in the short-scale convention. One millinillion occurs at $10^{6,000}$ instead of $10^{3,003}$. Finally, our last label convention would have a variation of Equation 6 that would use successive powers of $2*10^3$ vice 10^3.

But the limiting distribution would still involve only the function $P_0(\lambda)$, though $P_0(\lambda)$ for the long-scale convention would have all the extra "thousand" terms (or "illiard" terms if you prefer the alternative) in the labels. These additional "thousand" terms would affect our limiting distribution considerably. Computation of the long-scale variant of the limiting distribution is left for the reader.

We close by showing, in Figure 10, the cumulative distribution of the frequency of letters making up the names of the positive integers from 1 to ∞. Five letters ("i," "n," "t," "l," and "e") make up 60% of the letters in the distribution. The letter "i" has the largest frequency of about 16.7%. All letters after "b" have 0% frequency; "b" itself has a frequency of about 0.0044%.

Figure 10. Cumulative distribution of letter frequency in the limit



Appendix

The keys to reducing Equation 10 are the following formulae:

$$\sum_{i=1}^{M-1} x^i = \frac{1 - x^M}{1 - x} - 1$$

$$\sum_{i=1}^{M-1} i x^i = \frac{(M-1)x^{M+1} - Mx^M + x}{(1-x)^2}$$

From these, we find (setting i=j+1-M, x=10^{-3}, and ignoring terms on the right with M in the exponent and fixed terms)

$$\sum_{j=0}^{M-2} 10^{3(j+1-M)} = \frac{1 - 10^{-3M}}{1 - 10^{-3}} - 1 \approx \frac{1}{999}$$

and

$$\sum_{j=0}^{M-2} (j+1)10^{3(j+1-M)} = M\frac{1 - 10^{-3M}}{1 - 10^{-3}} - \frac{(M-1)10^{-3(M+1)} - M10^{-3M} + 10^{-3}}{(1-10^{-3})^2}$$

$$\approx \frac{M-1}{999} - \frac{1}{999^2} \approx \frac{M-1}{999}$$

Substituting the rightmost approximations for the sums into Equation 10 generates the approximation in Equation 11 for the cumulative average.

Acknowledgements

I want to thank Russ Beland (no longer of Springfield) for the puzzle suggestion that led to this paper. His question was somewhat simpler: in the names of the positive integers, how many "b"s do we see before we come across our first "c"? The "b"s, of course, first appear at one billion (10^9) and the first "c" appears at one octillion (10^{27}).

The exact answer to his question is $999 * 10^{24}$.

I want also to thank John Ivancovich for asking the "obvious" question of whether the labeling convention matters in the limiting case, which it does. I put "obvious" in quotes because it was "obvious" to me only after John asked it.

Finally, I thank my wife Terri for her patience as I spent a few weekends working out the math. She always reacted with good humor (perhaps bewildered amusement is more accurate) as I made headway on the problem.