ON THE SUMS OF DIGITS OF FIBONACCI NUMBERS

David C. Terr
Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720
(*Submitted December 1994*)

1. INTRODUCTION

The problem of determining which integers k are equal to the sum of the digits of F_k was first brought to my attention at the Fibonacci Conference in Pullman, Washington, this summer (1994). Professor Dan Fielder presented this as an open problem, having obtained all solutions for $k \le 2000$. There seemed to be fairly many solutions in base 10, and it was not clear whether there were infinitely many. Shortly after hearing the problem, it occurred to me why there were so many solutions. If one assumes that the digits F_k are independently uniformly randomly distributed, then one expects $S(k)$, the sum of the digits of F_k, to be approximately $\frac{9}{2}k\log_{10}\alpha$, where $\alpha = \frac{1}{2}(1+\sqrt{5}) \approx 1.61803$ is the golden mean. Since $\frac{9}{2}\log_{10}\alpha \approx 0.94044$, we expect $S(k) \approx 0.94044k$. Since this is close to k, we expect many solutions to $S(k) = k$, at least for reasonably small k. However, as k gets large, we expect $S(k)/k$ to deviate from 0.94044 by less and less. Thus, it appears that, for some integer n_0, the ratio $S(k)/k$ never gets as large as 1 for $k > n_0$, so $S(k) = k$ has no solutions for $k > n_0$, and thus has finitely many solutions. In this paper, I present two closely related probabilistic models to predict the number of solutions. More generally, they predict $N(b;n)$, the number of solutions to $S(k;b) = k$ for $k \le n$, where $S(k;b)$ is the sum of the digits of F_k in base b [thus, $S(k;10) = S(k)$]. Let $N(b)$ denote the total number of solutions in base b [thus, $N(b;n) \to N(b)$ as $n \to \infty$]. Both models predict finite values of $N(b)$ for each base b. In the simpler model, $N(10)$ is estimated to be 18.24 ± 3.86, compared with the actual value $N(10;20000) = 20$.

2. THE NAIVE MODEL

In this model I assume that the digits of F_k are independently uniformly randomly distributed among $\{0, 1, \ldots, b-1\}$ for each positive integer k and each fixed base $b \ge 4$. [It is fairly easy to prove that the only solutions to $S(k;b) = k$ are 0 and 1 when $b = 2$ or 3. The proof involves showing that, for all sufficiently large k, we have $(b-1)(1+\log_b F_k) < k$.] Now let $\alpha = \frac{1}{2}(1+\sqrt{5})$ and $\beta = \frac{1}{2}(1-\sqrt{5})$. Then

$$F_k = \frac{\alpha^k - \beta^k}{\sqrt{5}} = \frac{\alpha^k}{\sqrt{5}} + o(1) \text{ for } k \to \infty.$$

The number of digits of F_k in base b is approximately the base-b logarithm of this number, $k\log_b\alpha - \log_b\sqrt{5} \approx k\log_b\alpha = k\gamma$, where $\gamma = \log_b\alpha$ and I neglect terms of order 1. In this model, the expected value of each digit of F_k is $\frac{1}{2}(b-1)$ and the standard deviation (SD) is $\sqrt{\frac{1}{12}(b^2-1)}$ (see [2], pp. 80-86). Therefore, the expected value of $S(k;b)$ is approximately $\overline{S} = \frac{k}{2}(b-1)\gamma$ and the SD is approximately $\sigma = \sqrt{\frac{k}{12}(b^2-1)\gamma}$. Let $\mathcal{P}_1(k;\ell)$ denote the probability that $\tilde{S}(k;b) = \ell$, where $\tilde{S}(k;b)$ is distributed as the sum $Y_{k,1} + Y_{k,2} + \cdots + Y_{k,[k\gamma]}$, the $Y_{k,j}$ being independent random variables, each uniformly distributed over $\{0, 1, \ldots, b-1\}$. According to the central limit theorem ([2], pp. 165-77), if k is reasonably large, the probability distribution is approximately Gaussian, so

$$\mathcal{P}_1(k;\ell)\approx\frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(\overline{S}-\ell)^2}{2\sigma^2}\right]=\sqrt{\frac{6}{k\pi\gamma(b^2-1)}}\exp\left[\frac{-6\left(k\gamma\left(\frac{b-1}{2}\right)-\ell\right)^2}{k\gamma(b^2-1)}\right].$$

Let $\mathcal{P}_1(k)=\mathcal{P}_1(k;k)$; this is the estimated percentage of Fibonacci numbers $F_{k'}$ for k' near k whose base-b digits sum to the index k'. We have $\mathcal{P}_1(k)\approx Ae^{-Bk}/\sqrt{k}$, where

$$A=\sqrt{\frac{6}{\pi\gamma(b^2-1)}}\quad\text{and}\quad B=\frac{6\left(\gamma\left(\frac{b-1}{2}\right)-1\right)^2}{\gamma(b^2-1)}.$$

Incidentally, it is clear that the only solutions k for which $F_k<b$ are those for which $F_k=k$, namely 0, 1, and possibly 5 (if $b>5$). We might as well put in these solutions by hand. Thus, in the model, we only calculate $\mathcal{P}_1(k)$ for k for which $F_k\geq b$ and add N_0 to the final result upon summing the probabilities, where $N_0=3$ if $b>5$, otherwise $N_0=2$. Thus, our estimate for $N(b;n)$ in this model is

$$N_1(b;n)=N_0+\sum_{\substack{k\leq n\\F_k\geq b}}\mathcal{P}_1(k)\approx N_0+\sum_{\substack{k\leq n\\F_k\geq b}}\frac{Ae^{-Bk}}{\sqrt{k}}$$

and the standard deviation of this estimate is [assuming that the $S(k,b)$ are uncorrelated for different values of k]

$$\Delta_1(b;n)=\sqrt{\sum_{\substack{k\leq n\\F_k\geq b}}\mathcal{P}_1(k)(1-\mathcal{P}_1(k))}\approx\sqrt{\sum_{\substack{k\leq n\\F_k\geq b}}\frac{Ae^{-Bk}(\sqrt{k}-Ae^{-Bk})}{k}}.$$

This model gives good results for some bases, but not all. The next model is an improvement which seems to yield accurate results for all bases.

3. THE IMPROVED MODEL

In this model, I still assume that the digits of F_n are uniformly distributed over $\{0,1,\ldots,b-1\}$, but with one restriction, namely, their sum modulo $b-1$. It is well known that the sum of the base-10 digits of a number a is congruent to a mod 9. In general, the same applies to the sum of the digits in base b modulo $b-1$. Thus, we have the restriction $S(k;b)\equiv F_k\ (\operatorname{mod} b-1)$. In particular, k cannot be a solution to $S(k;b)=k$ unless $k\equiv F_k\ (\operatorname{mod} b-1)$. This latter equation is not too difficult to solve. Upon solving it, we end up with a restriction of the form

$$k \bmod q\in S. \tag{1}$$

Here, $q=[b-1,p]$, where $p=\operatorname{per}(b-1)$ is the period of the Fibonacci sequence modulo $b-1$ and S is a specified subset of $\{0,1,\ldots,q-1\}$. If k does not satisfy the above condition, it need not be considered, since the sum of its digits cannot equal F_n. On the other hand, if k does satisfy the condition, we know that the sum of its digits is congruent to F_n modulo $b-1$. In the improved model, we take this restriction into account and otherwise assume a uniform random distribution of digits in F_n. In analogy to $\tilde{S}(k;b)$, let $\hat{S}(k;b)$ be distributed as the sum $Y_{k,1}+\cdots+Y_{k,[k\gamma]}$, the $Y_{k,j}$ being random variables uniformly distributed over 0, 1, ..., $b-1$ and independent except for the restriction that $Y_{k,1}+\cdots+Y_{k,[k\gamma]}\equiv F_k \pmod{b-1}$. We now estimate the probability $\mathcal{P}_2(k)$ that $\hat{S}(k;b)=k$ to be $b-1$ times our earlier estimate in the case where k satisfies (1) and zero otherwise, i.e.,

$$\mathcal{P}_2(k)=\begin{cases}(b-1)\mathcal{P}_1(k) & k \bmod q\in S,\\ 0 & k \bmod q\notin S.\end{cases}$$

Thus, in this model, the expectation and SD of $N(b;n)$ are approximately

$$N_2(b;n)=N_0+\sum_{\substack{k\le n\\ F_k\ge b}}\mathcal{P}_2(k)\approx N_0+(b-1)\sum_{\substack{k\le n\\ F_k\ge b\\ k \bmod q\in S}}\frac{Ae^{-Bk}}{\sqrt{k}}$$

and

$$\Delta_2(b;n)=\sqrt{\sum_{\substack{k\le n\\ F_k\ge b}}\mathcal{P}_2(k)(1-\mathcal{P}_2(k))}\approx\sqrt{(b-1)\sum_{\substack{k\le n\\ F_k\ge b\\ k \bmod q\in S}}\frac{Ae^{-Bk}(\sqrt{k}-Ae^{-Bk})}{k}}.$$

As an example of how to calculate S, consider $b=8$. In this case, $p=\text{per}(7)=16$ and $q=[7,16]=112$. To determine S, we first tabulate k mod 16 and F_k mod 7 for each congruence class of k mod 16. Next, below the line, we tabulate the unique solutions modulo 112 to the congruences $x\equiv k$ (mod 16) and $x\equiv F_k$ (mod 7). Since $(16,7)=1$, by the Chinese Remainder Theorem, each of these solutions exists and is unique.

k mod 16	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F_k mod 7	0	1	1	2	3	5	1	6	0	6	6	5	4	2	6	1
k mod 112	0	1	50	51	52	5	22	55	56	41	90	75	60	93	62	15

Thus, $S=\{0,1,5,15,22,41,50,51,52,55,56,60,62,75,90,93\}$. Note that, in this example, the pair of congruences $k\equiv j \pmod p$ and $k\equiv F_j$ mod $b-1$ has a solution mod q for every integer j mod p. This is because , in this example, $b-1=7$ and $p=16$ are coprime. In general, this is not the case. For example, for $b=10$, we get $p=24$, which is not coprime to $b-1=9$. Thus, if we constructed a similar table for $b=10$, we would expect to get some simultaneous congruences without solutions. This is in fact the case, i.e., the pair of congruences $k\equiv 2$ (mod 24) and $k\equiv F_2=1$ (mod 9) has no solutions. We expect only one-third (eight) of them to have a solution, since $(9,24)=3$. In fact, we do get eight. For $b=10$, we find $S=\{0,1,5,10,31,35,36,62\}$ and $q=72$.

One might wonder by about how much $N_1(k;b)$ and $N_2(k;b)$ differ. To first order, they differ by a multiplicative factor depending on b, i.e., $N_2(b;n)\approx M(b)N_1(b;n)$. Recall that in going from the first model to the second, we selected s out of every q congruence classes modulo q, where $s=\#S$. Also, we multiplied the corresponding probabilities by $b-1$. Thus, $M(b)=(b-1)s/q$. For some bases, $M(b)=1$, so the predictions of both models are essentially the same. This is true in particular whenever $b-1$ and p are coprime, and also in some other cases, like $b=10$. However, there are other bases for which $M(b)\neq 1$; in fact, the difference can be quite large! For instance, for $b=11$, we find $p=q=60$ and $s=14$, hence $M(11)=10\times 14/60=7/3$, which is greater than 2. Thus, for $b=11$, the second model predicts over twice as many solutions to $S(k;11)=k$ as the first model. In this case, as we will see, the second model agrees well with the known data; the first does not.

4. COMPARISON OF MODELS WITH "EXPERIMENT"

Every good scientist knows that the best way to test a model or theory is to see how well its predictions agree with experimental data. In this case, my "experiment" was a computer program I wrote and ran on my Macintosh LCII to determine $S(k;b)$ given $k\le 20000$ and $b\le 20$. Incidentally, it is not necessary to calculate the Fibonacci numbers directly, only to store the digits in an array. Also, only two Fibonacci arrays need to be stored at one time. Nevertheless, trying to compute for $k>20000$ presented memory problems, at least for the method I used. Still, this turned out to be sufficient for determining with high certainty all solutions to $S(k;b)=k$ except for $b=11$.

Here I present all the solutions I found for $4\le b\le 20$ and $k\le n$.

b=4, n=1000:	0	1						
b=5, n=1000:	0	1						
b=6, n=1000:	0	1	5	9	15	35		
b=7, n=1000:	0	1	5	7	11	12	53	
b=8, n=1000:	0	1	5	22	41			
b=9, n=5000:	0	1	5	29	77	149	312	
b=10, n=20000:	0	1	5	10	31	35	62	72
	175	180	216	251	252	360	494	504
	540	946	1188	2222				
b=11, n=20000:	0	1	5	13	41	53	55	60
	61	90	97	169	185	193	215	265
	269	353	355	385	397	437	481	493
	617	629	630	653	713	750	769	780
	889	905	960	1013	1025	1045	1205	1320
	1405	1435	1501	1620	1650	1657	1705	1735
	1769	1793	1913	1981	2125	2153	2280	2297
	2389	2413	2460	2465	2509	2533	2549	2609
	2610	2633	2730	2749	2845	2893	2915	3041
	3055	3155	3209	3360	3475	3485	3521	3641
	3721	3749	3757	3761	3840	3865	3929	3941
	4075	4273	4301	4650	4937	5195	5209	5435
	5489	5490	5700	5917	6169	6253	6335	6361
	6373	6401	6581	6593	6701	6750	6941	7021
	7349	7577	7595	7693	7740	7805	7873	8009
	8017	8215	8341	8495	8737	8861	8970	8995
	9120	9133	9181	9269	9277	9535	9541	9737
	9935	9953	10297	10609	10789	10855	11317	11809
	12029	12175	12353	12461	12565	12805	12893	13855
	14381	14550	14935	15055	15115	15289	15637	15709
	16177	16789	16837	17065	17237	17605	17681	17873
	17941	17993	18193	18257	18421	18515	18733	18865
	18990	19135	19140	19375	19453	19657	19873	

b=12, n=20000:	0	1	5	13	14	89	96	123
	221	387	419	550	648	749	866	892
	1105	2037						
b=13, n=5000:	0	1	5	12	24	25	36	48
	53	72	73	132	156	173	197	437
	444	485	696	769	773			
b=14, n=3000:	0	1	5	8	11	27	34	181
	192	194						
b=15, n=2000:	0	1	5					
b=16, n=2000:	0	1	5	10	60	101		
b=17, n=1000:	0	1	5					
b=18, n=1000:	0	1	5	60				
b=19, n=1000:	0	1	5	31	36			
b=20, n=1000:	0	1	5	21	22			

Next, I tabulated $N_1(b;n)\pm\Delta_1(b;n)$, $N_2(b;n)\pm\Delta_2(b;n)$, $N(b;n)$, $(\frac{b-1}{2})\log_b\alpha$, and $M(b)$ for the above pairs (b,n). Note how $N(b;n)$ increases as $(\frac{b-1}{2})\log_b\alpha$ approaches 1.

b	n	$N_1(b;n)\pm\Delta_1(b;n)$	$N_2(b;n)\pm\Delta_2(b;n)$	$N(b;n)$	$(\frac{b-1}{2})\log_b\alpha$	$M(b)$
4	1000	2.25 ± 0.49	2.43 ± 0.61	2	0.52068	1.00
5	1000	2.67 ± 0.79	2.76 ± 0.72	2	0.59799	1.00
6	1000	4.17 ± 1.05	5.04 ± 1.25	6	0.67142	2.00
7	1000	5.10 ± 1.41	6.18 ± 1.42	7	0.74188	1.50
8	1000	6.61 ± 1.85	5.84 ± 1.47	5	0.80995	1.00
9	5000	9.40 ± 2.48	8.57 ± 2.09	7	0.87604	1.00
10	20000	18.24 ± 3.86	17.77 ± 3.46	20	0.94044	1.00
11	20000	79.71 ± 8.72	180.95 ± 12.82	183	1.00340	2.33
12	20000	17.03 ± 3.71	17.01 ± 3.28	18	1.06510	1.00
13	5000	9.71 ± 2.56	15.73 ± 3.08	21	1.12566	2.00
14	3000	7.15 ± 2.01	8.22 ± 1.62	10	1.18522	1.00
15	2000	5.93 ± 1.69	4.70 ± 1.19	3	1.24387	1.00
16	2000	5.21 ± 1.47	7.16 ± 1.62	6	1.30170	2.00
17	1000	4.75 ± 1.31	3.94 ± 0.90	3	1.35877	1.00
18	1000	4.42 ± 1.18	4.69 ± 1.06	4	1.41515	1.00
19	1000	4.10 ± 1.08	4.12 ± 0.95	5	1.47088	1.50
20	1000	4.01 ± 1.00	4.54 ± 0.97	5	1.52601	1.00

As one can see, the first model does not make accurate predictions for each base. In particular, its predictions for bases 11 and 13 are off by roughly 12 and 4.5 standard deviations, respectively. On the other hand, the second model seems to agree well with the known data for each base. For 12 out of 17 bases, its predictions are correct within one SD, and all 17 predictions are correct within two SD's. (The largest deviation, found for $b=13$, is -1.71 SD's.) Furthermore, there does not seem to be a directional bias of the model. Eight out of 17 of the predicted values are too high; the other 9 are too low. Thus, the second model looks good.

5. PREDICTING THE UNKNOWN

With this in mind, we can use the second model to make predictions for which we are unable to calculate at present. In particular, we can estimate $N(11)$, the total number of solutions to $S(k;11)=k$ in base 11, as well as the value of the largest one. We can also estimate the probability that we missed some solutions in each of the other bases we looked at. For these bases, I was careful to calculate out to large enough n so that these probabilities should be very small.

I calculated $N_2(11;n)\pm\Delta_2(11;n)$ for $200000\le n\le 4000000$ in intervals of 200000. Here are the results:

n	$N_2(11;n)\pm\Delta_2(11;n)$
200000	490.38 ± 21.70
400000	595.89 ± 24.00
600000	641.02 ± 24.93
800000	662.32 ± 25.35
1000000	672.83 ± 25.56
1200000	678.16 ± 25.66
1400000	680.91 ± 25.71
1600000	682.34 ± 25.74
1800000	683.10 ± 25.76
2000000	683.50 ± 25.77
2200000	683.72 ± 25.77
2400000	683.83 ± 25.77
2600000	683.89 ± 25.77
2800000	683.93 ± 25.77
3000000	683.94 ± 25.77
3200000	683.95 ± 25.77
3400000	683.96 ± 25.77
3600000	683.96 ± 25.77
3800000	683.96 ± 25.77
4000000	683.97 ± 25.77

As can be seen, the results converge rapidly for large n. Let $N'(b;n)$ denote the estimated number of solutions to $S(k;b)=k$ for $k>n$. Then we have

$$N'(b;n)=\sum_{\substack{k>n\\ k \bmod q\in S}}\frac{Ae^{-Bk}}{\sqrt{k}}\approx M(b)\sum_{k>n}\frac{Ae^{-Bk}}{\sqrt{k}}\approx\frac{M(b)A}{\sqrt{B}}\int_{Bn}^{\infty}\frac{e^{-x}dx}{\sqrt{x}}$$

$$\approx M(b)A\sqrt{\frac{\pi}{B}}\operatorname{erfc}\sqrt{Bn}\approx\frac{M(b)A}{B\sqrt{n}}e^{-Bn},$$

where I make the change of variables $y=\sqrt{x}$ in the integral to get the error function term. In the last step, I use an asymptotic expansion of erfc [1].

I next tabulated $N'(b;n)$ for the pairs (b,n) used, except for $b=11$, where I used $n=4000000$, the largest n for which I have estimated $N(b;n)$. Since N' is much less than 1 in each case, the values of N' listed are the approximate probabilities that there is a solution to $S(k;b)=k$ for $k>n$. I also tabulated the corresponding values of A, B, and $M(b)$. Note that for every base less than 20, except 11, $N'(b;n)$ is less than 10^{-6}; in fact, the sums of these entries is roughly 10^{-6}. Thus, if this model is accurate, there is about one chance in a million that I have missed any solutions in these bases. Also, note that the table of estimates of $N_2(11;n)$ can be used to estimate the largest solution to $S(k;11)=k$.

b	n	A	B	$M(b)$	$N'(b;n)$
4	1000	0.606	2.65×10^{-1}	1.000	7.6×10^{-117}
5	1000	0.516	1.35×10^{-1}	1.000	2.5×10^{-60}
6	1000	0.451	6.89×10^{-2}	2.000	4.7×10^{-31}
7	1000	0.401	3.37×10^{-2}	1.500	1.3×10^{-15}
8	1000	0.362	1.49×10^{-2}	1.000	2.7×10^{-7}
9	5000	0.330	5.26×10^{-3}	1.000	2.3×10^{-12}
10	20000	0.304	1.03×10^{-3}	1.000	2.4×10^{-9}
11	4000000	0.282	2.89×10^{-6}	2.333	1.1×10^{-3}
12	20000	0.263	9.18×10^{-4}	1.000	2.1×10^{-9}
13	5000	0.246	3.01×10^{-3}	2.000	6.8×10^{-7}
14	3000	0.232	5.79×10^{-3}	1.000	1.6×10^{-8}
15	2000	0.219	8.97×10^{-3}	1.000	7.2×10^{-9}
16	2000	0.208	1.23×10^{-2}	2.000	1.7×10^{-11}
17	1000	0.198	1.58×10^{-2}	1.000	5.5×10^{-8}
18	1000	0.188	1.92×10^{-2}	1.000	1.4×10^{-9}
19	1000	0.180	2.26×10^{-2}	2.000	5.2×10^{-11}
20	1000	0.172	2.59×10^{-2}	1.000	1.1×10^{-12}

Suppose one wishes to find n such that there is a 50% chance that there are no solutions larger than n. According to Poisson statistics, this happens when the $N_2(11; n) = \ln 2 \approx 0.69$. By interpolating in the previous table, we see that this occurs when $n \approx 1.9 \times 10^6$; this is roughly the value we can expect for the largest solution. Calculating $S(k; 11)$ for k up to 2.8×10^6 yields a 96% probability of finding all the solutions, and going up to 4×10^6 yields a 99.9% probability of finding them all. Perhaps someone will do this calculation in the near future.

ACKNOWLEDGMENTS

I would like to give special thanks to my advisor, H. W. Lenstra, Jr., for helping me with this paper. In particular, he made me aware of the need for congruence relations in finding solutions to $S(k; b) = k$.

REFERENCES

1. M. Abramowitz & I. A. Stegun. *Handbook of Mathematical Functions.* 9th ed. New York: Dover Publications, 1970.
2. H. J. Larson. *Statistics: An Introduction.* New York, London: Wiley, 1975.

AMS Classification Numbers: 11A63, 11B39, 60F99

❖❖❖