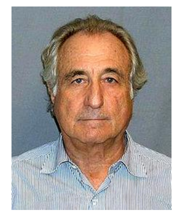
Benford's Law

Benford's Law predicts the occurrence of digits in large sets of data. Simply put, this law maintains that we can expect some digits to occur more often than others. For example, the numeral 1 should occur as the first digit in any multiple-digit number about 30.1% of the time, while the numeral 9 should occur as the first digit only 4.6% of the time. We also can apply the law to determine the expected occurrence of the second digit of a number, the first two digits of a number and other combinations.

How can such predictions help you red-flag potential irregularities? When someone creates false transactions or commits a data-entry error, the resulting numbers often deviate from the law's expectations. This is true when someone creates random numbers or intentionally keeps certain transactions below required authorization levels.

For example, in 2008, Bernie Madoff famously created fictitious data to hide an estimated $65 billion in losses resulting from Madoff's investment Ponzi scheme. – Had the CPAs looked at this data using Benford's Law, they might have found that the digits smelled of fraud, perhaps triggering a deeper investigation.



Applying Benford's Law Using Excel

According to Benford's Law, the various digits should occur as the first digit position according to the following percentages.

Benford's Law

Number	Occurrence As First Digit
1	30.10%
2	17.60%
3	12.50%
4	9.70%
5	7.90%
6	6.70%
7	5.80%
8	5.10%
9	4.60%

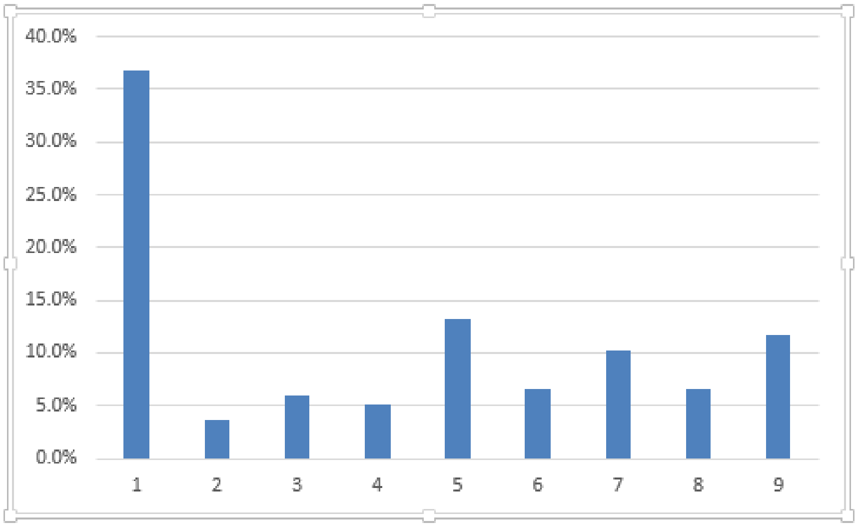
To analyze data, simply use the LEFT function to extract the leading digits, and then add them up as follows. As a simple example, I found a random workbook containing 136 rows of revenue amounts. I entered a formula in cell C4 to extract the first digit and copied this formula down. Next I used the

COUNTIF function to count the number of occurrences of each of the nine digits, and calculated their rate of occurrence, then charted the results.

C4 =LEFT(A4,1)

	A	C
3	**Revenue**	
4	**47,520**	**4**
5	**91,463**	**9**
6	**356,040**	**3**
7	**154,276**	**1**
8	**113,924**	**1**
9	**94,936**	**9**
10	**90,305**	**9**
11	**394,276**	**3**
12	**116,122**	**1**
13	**92,621**	**9**
14	**69,466**	**6**
15	**261,005**	**2**
16	**57,024**	**5**
17	**181,988**	**1**
18	**121,198**	**1**
19	**432,900**	**4**
20	**96,768**	**9**
21	**121,198**	**1**
22	**121,198**	**1**
23	**110,772**	**1**
24	**128,563**	**1**
25	**128,563**	**1**
26	**76,032**	**7**
27	**110,772**	**1**

E	F	G
1	**50**	**36.8%**
2	**5**	**3.7%**
3	**8**	**5.9%**
4	**7**	**5.1%**
5	**18**	**13.2%**
6	**9**	**6.6%**
7	**14**	**10.3%**
8	**9**	**6.6%**
9	**16**	**11.8%**
	136	**100.0%**



You can clearly see that this data pattern does not conform to Benford's law, and yes, I fabricated this particular set of data years ago.

When Excel helps you spot a deviation like this, it raises a red flag. Considerable statistical research supports the effectiveness of Benford's Law, making it a valuable tool for CPAs. The technique isn't guaranteed to detect fraud in all situations but is useful in analyzing the credibility of accounting records.

A Note of Caution

Benford's Law is not effective for all financial data. If the data set is small, the law becomes less accurate because there are not enough items in the sample and so the rules of randomness don't apply—or at least apply with less predictability.

Also, if the data include built-in minimums and maximums, they also might not conform well to the law's predictions. For example, consider a petty-cash fund where all disbursements are between a $10 minimum and a $20 maximum. All first digits would be either 1 or 2, and the expected distribution of first digits would not apply. Likewise, when a company's major product sells for, say, $9.95, most sales totals will be a multiple of 995, again offsetting the value of the process. Finally, when a data set consists of assigned numbers, such as a series of internally generated invoice numbers, the data will not follow a Benford distribution.

History of Beford's Law

Newcomb, 1881 - The discovery of Benford's law dates back to 1881, when the American astronomer Simon Newcomb noticed that in logarithm tables (used at that time to perform calculations) the earlier pages (which contained numbers that started with 1) were much more worn than the other pages. Newcomb's published result is the first known instance of this observation and includes a distribution on the second digit, as well. Newcomb proposed a law that the probability of a single number N being the first digit of a number was equal to $\log(N + 1) - \log(N)$.

Benford, 1938 - The phenomenon was again noted in 1938 by the physicist Frank Benford, who tested it on data from 20 different domains and was credited for it. His data set included the surface areas of 335 rivers, the sizes of 3259 US populations, 104 physical constants, 1800 molecular weights, 5000 entries from a mathematical handbook, 308 numbers contained in an issue of Readers' Digest, the street addresses of the first 342 persons listed in American Men of Science and 418 death rates. The total number of observations used in the paper was 20,229. This discovery was later named after Benford.

Varian, 1972 - In 1972, Hal Varian suggested that the law could be used to detect possible fraud in lists of socio-economic data submitted in support of public planning decisions. Based on the plausible assumption that people who make up figures tend to distribute their digits fairly uniformly, a simple comparison of first-digit frequency distribution from the data with the expected distribution according to Benford's law ought to show up any anomalous results.

Nigrini, 1999 - Following this idea, Mark Nigrini showed that Benford's law could be used in forensic accounting and auditing as an indicator of accounting and expenses fraud. In practice, applications of Benford's law for fraud detection routinely use more than the first digit.

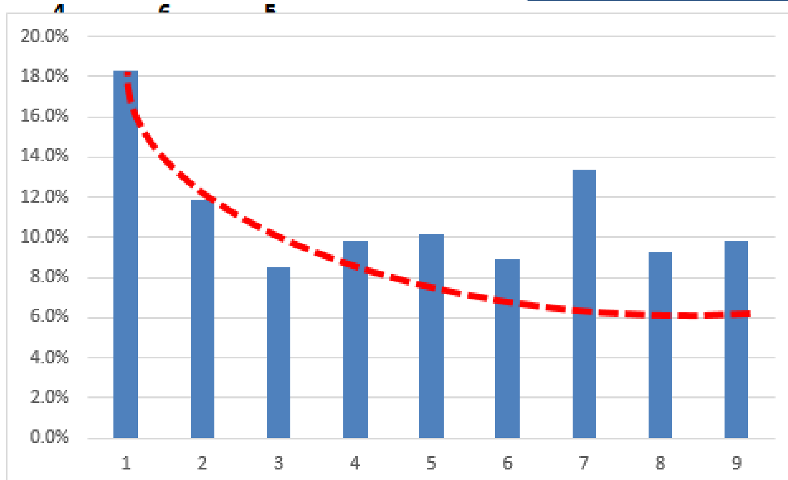
Analyzing All Digits

Building on the simple example described above, we can expand our Excel formulas to analyze all of the digits included in a set of data as follows: In the example shown below, I have used the MID function to extract each digit from the column of values in column **A**, and the resulting individual numerals are displayed in columns **C** through **H**. Next, I totaled the occurrence for each number 1 through 9 in the summary box and charted the results.

E4 | =MID($A4,E$3,1)

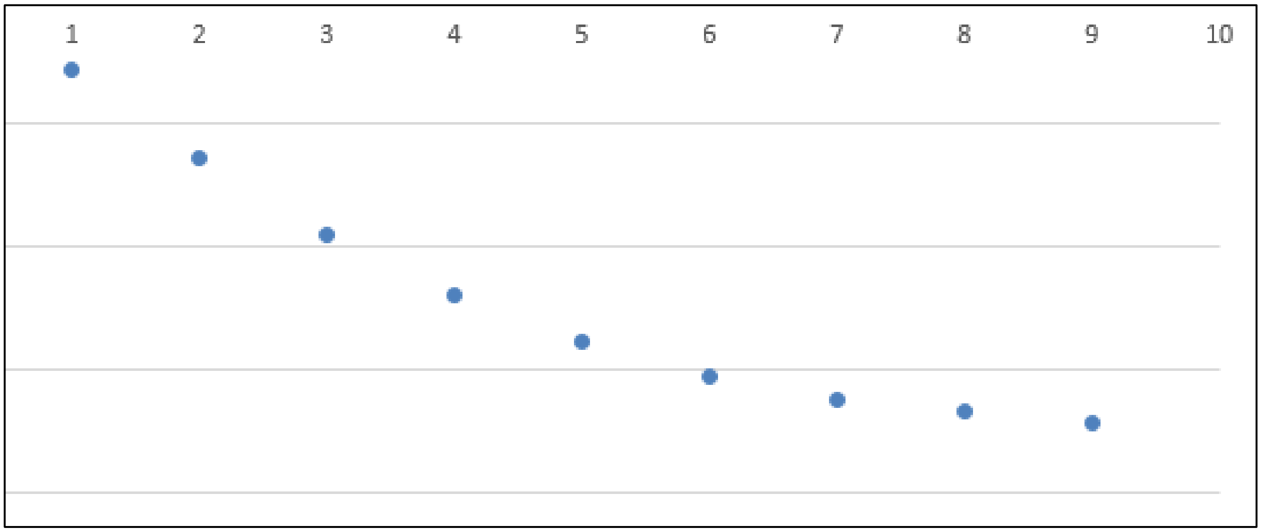
	A	B	C	D	E	F	G	H
3	**Revenue**		1	2	3	4	5	6
4	47,520		4	7	5	2	0	
5	91,463		9	1	4	6	3	.
6	356,040		3	5	6	0	4	0
7	154,276		1	5	4	2	7	5
8	113,924		1	1	3	9	2	3
9	94,936		9	4	9	3	6	.
10	90,305		9	0	3	0	5	.
11	394,276		3	9	4	2	7	5
12	116,122		1	1	6	1	2	1
13	92,621		9	2	6	2	0	.
14	69,466		6	9				
15	261,005		2	6				
16	57,024		5	7				
17	181,988		1	8				
18	121,198		1	2				
19	432,900		4	3				
20	96,768		9	6				
21	121,198		1	2				
22	121,198		1	2				
23	110,772		1	1				
24	128,563		1	2				
25	128,563		1	2				
26	76,032		7	6				
27	110,772		1	1				
28	110,772		1	1	0	7	7	2
29	91,224		9	1	2	2	4	

J	K	L
1	123	18.3%
2	80	11.9%
3	57	8.5%
4	66	9.8%
5	68	10.1%
6	60	8.9%
7	90	13.4%
8	62	9.2%
9	66	9.8%
	672	100.0%



In this example, the data does appear to ever so slightly adhere to Benford's Law as the first 4 bars in the chart and a few others seem to come close to matching Benford's declining curve.

According to Benford's Law one would expect lower numerals to appear more frequently than higher values, but why? Lower digits (1, 2,& 3) tend to appear more frequently than higher digits (7, 8, & 9) because it is easier to own 1 acre than 9 acres; and more people have $100 than $900. Lower numbers are typically more achievable than higher numbers in many situations. For this reason, we would expect to see an analysis of numbers form a slightly curved declining chart like this one:

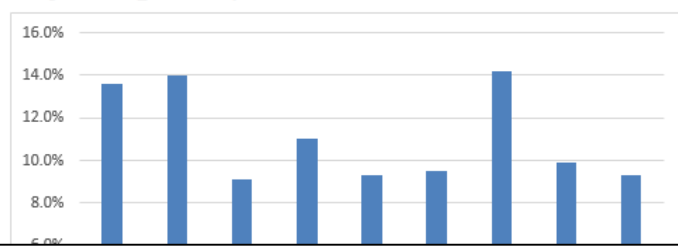


However, the presence of the leading digit may significantly skew the data, therefore, we could ignore the leading digit and analyze the occurrence of the remaining 8 numerals in an effort to determine whether or not the data appears to roughly follow Benford's Law. Ignoring the leading digit reveals the following analysis.

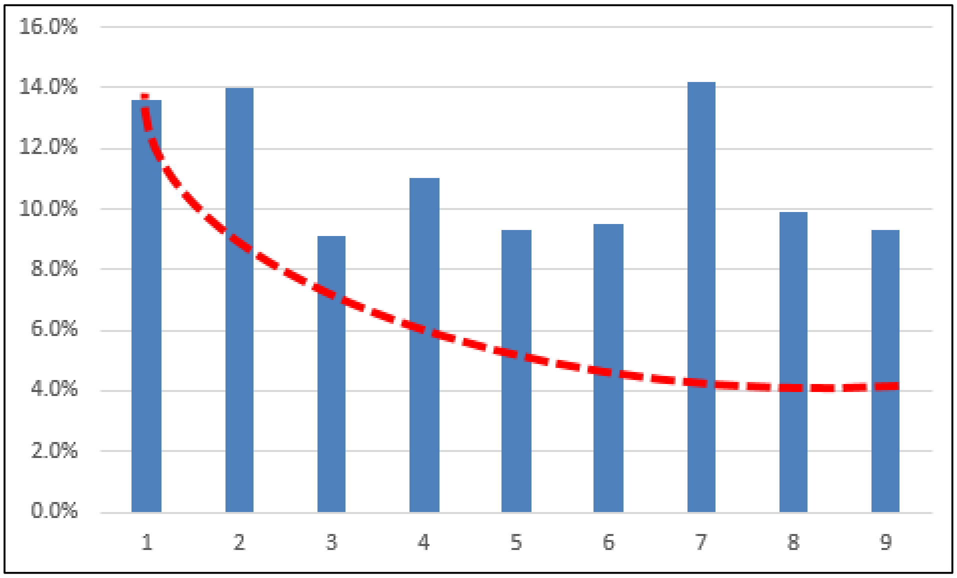
E4 =MID($A4,E$3,1)

	A	B	C	D	E	F	G	H
1								
2								
3	Revenue		1	2	3	4	5	6
4	47,520			7	5	2	0	
5	91,463			1	4	6	3	.
6	356,040			5	6	0	4	0
7	154,276			5	4	2	7	5
8	113,924			1	3	9	2	3
9	94,936			4	9	3	6	.
10	90,305			0	3	0	5	.
11	394,276			9	4	2	7	5
12	116,122			1	6	1	2	1
13	92,621			2	6	2	0	.
14	69,466			9	4	6	5	.
15	261,005			6	1	0	0	4
16	57,024			7	0	2	4	
17	181,988			8				
18	121,198			2				
19	432,900			3				
20	96,768			6				
21	121,198			2				
22	121,198			2				
23	110,772			1				
24	128,563			2				

J	K	L
1	73	13.6%
2	75	14.0%
3	49	9.1%
4	59	11.0%
5	50	9.3%
6	51	9.5%
7	76	14.2%
8	53	9.9%
9	50	9.3%
	536	100.0%



As expected, the numeral 1 occurs less often than when the leading digits were included. Still, with the leading digit ignored, the data appears not to follow Benford's Law.



Conclusion

This case study was intentionally brief and is only intended to convey the general ideas related to Benford's law. The forensic CPA may choose to run these numbers to help confirm suspicions or beliefs related to the authenticity of a large data set of numbers in question. There are probably many instances where this approach would offer little value, however in a high ticket audit with high potential for fraud, running Benford's Law analysis is a rather quick exercise which may offer insights to help the forensic CPA determine how to best proceed.