European Commission Benford's Law Conference
Stresa, Italy July 10-12 2019

Benford's Law and Detection of Anomalies in Data

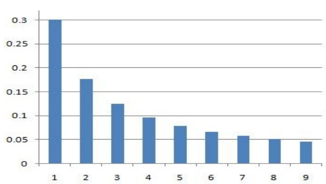
1 5 7 3 9 4 6 8 2

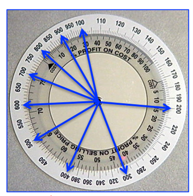
Dr. Ted Hill
School of Mathematics, Georgia Tech
California Polytechnic State University

Outline

- **Brief History of Benford's Law (BL)**
- **Use of BL to Detect Anomalies in Data**
  - **Fraud**
  - **Other anomalies**
- **Seven Basic BL Probability Theorems**
- **Common Errors related to BL**
- **How to win € from your friends**

Benford's Law for First Digits





$$\text{Prob}\big(\text{First digit of } X \text{ is } d\big) = \log_{10}\left(1+d^{-1}\right),\ d = 1,2,\ldots,9$$

i.e., $P(D_1(X)=1) = \log_{10}(2) \cong .301$

$P(D_1(X)=2) = \log_{10}(1.5) \cong .176$

...

$P(D_1(X)=9) = \log_{10}(1+0.111...) \cong .046$

(Here D_1 is the first significant digit (base 10) of $x > 0$.
e.g., $D_1(2019) = D_1(0.02019) = 2$)

40 NEWCOMB: *Note on the Frequency of Use of the Different Digits, Etc.*

We have a series of numbers between 1 and i, represented by fractional powers of i, say i^s, the distribution of the exponents s, and therefore of the numbers, being according to any arbitrary law. Since these exponents are formed by casting off all the integers from a series of numbers, we may suppose them arranged around a circle according to some law. Then, if we select 2^n exponents at random and call them s', s'', s''', etc., the final ratio, obtained in the manner we have described, will be

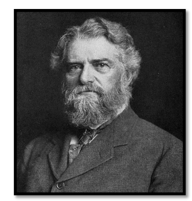
$$i^{s'-s''+s'''-s''''+\text{etc.}}$$

The question is, what is the probability that the positive fractional portion of $s' - s'' + s''' - s'''' +$ etc., will be contained between the limits s and $s + ds$. It is evident that, whatever be the original law of arrangement, the fractions will approach to an equal distribution around the circle as n is increased, or the required probability will be equal to ds. But, the fractional part of $s' - s'' + s''' -$ etc. is the mantissa of the logarithm of the limiting ratio. We thus reach the conclusion:

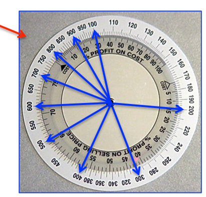
The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.

In other words, every part of a table of anti-logarithms is entered with equal frequency. We thus find the required probabilities of occurrence in the case of the first two significant digits of a natural number to be:

Dig.	First Digit.	Second Digit.
0	. . .	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850



Newcomb 1881



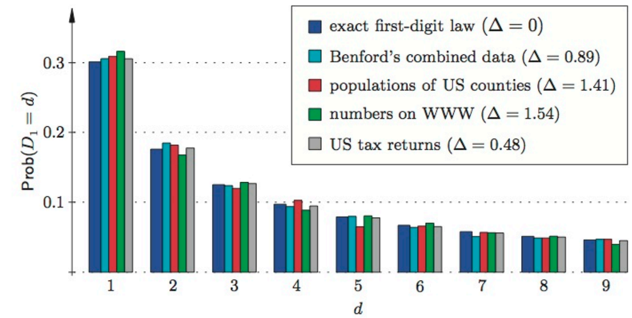
First-digit Dataset (Benford 1938)

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit 1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Empirical Evidence of BL Today



BL Fraud Detection

(Key Idea by **M. Nigrini** 1990's)

Tax (individual, corporate, governmental)

Clinical and drug trials

Survey data

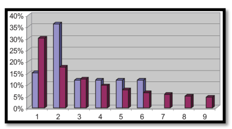
Environmental

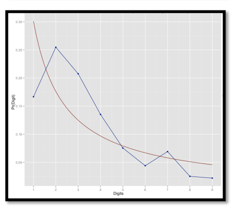
Voting

Health Insurance

Scientific papers

Fingerprint forgeries





Benford's Own Data?

Group	Title	First Digit 1	2	3	4	5	6	7	8	9	Count
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335

Diaconis&Freedman 1979:

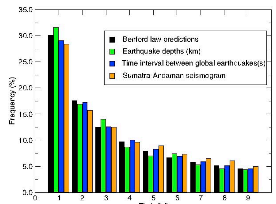
$\frac{18}{335}$ **rounds to 5.4% and** $\frac{19}{335}$ **rounds to 5.7%**

D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
BL. % :		**30.1**	**17.6**	**12.5**	**9.7**	**7.9**	**6.7**	**5.8**	**5.1**	**4.6**	

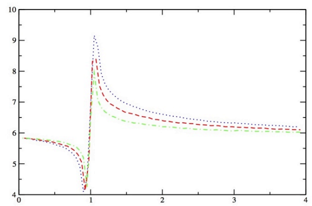
BL Other Anomaly Detection – Phase Transitions

Earthquakes (depths, time intervals)

Quantum processes (many-body problems)



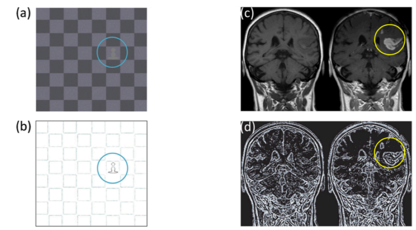
(Sambridge et al 2011)



(Sen and Sen 2011)

BL Other Anomaly Detection - Image Processing

Spectroscopic analysis (e.g., MRI's)



Steganography (hidden images)

Natural vs. artificial images

Image alterations

BL Other Anomaly Detection

Internet traffic
(intrusions, intentional & not)

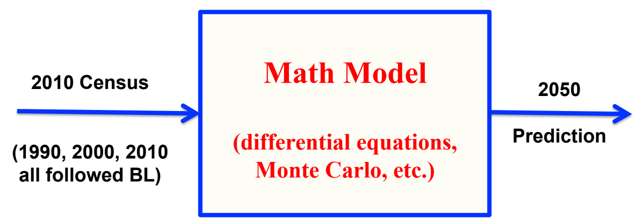
Music analysis
(natural vs. artificially created chords)

Sport game manipulation
(detect match-fixing)

Macroeconomics
(GDP, purchasing power parity)

Cardiology
(different types of arrhythmia)

Related Application – Model Testing



Benford-In, Benford-Out Test

Seven Basic BL Probability Theorems

Thm 1. *BL* is the unique **scale-invariant** probability distribution on significant digits.

Ex. If a financial dataset X is Benford in **€** it is also B in **$**.
If X is **not** Benford in **€** it is also **not** Benford in **$**

Ex. If distances to galaxies in light years follow BL, they will also follow BL measured in inches, centimeters, miles, and every other unit.

Thm 2. *BL* is the unique continuous **base-invariant** probability distribution on significant digits.

Thm 3. *BL* is the unique **sum-invariant** probability distribution on significant digits **(Nigrini, Allaart).**

BL Probability Theorems (cont'd)

Thm 4. If X is a Benford random variable, then so are
X^2, $1/X$, and XY,
where Y is any positive random variable independent of X.

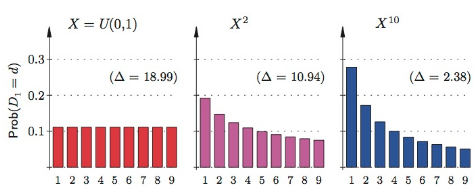
Ex. If a financial dataset X is Benford in **€ per stock**, it is also Benford in **stock per €**.

Ex. If $X_1 \times X_2 \times X_3 \times X_4 \times \ldots \times X_n$ are independent positive random variables (e.g. interest rates), then **if any X_i is Benford**, then the whole product is Benford and remains Benford forever.

BL Probability Theorems (cont'd)

Thm 5. If X is a random variable with a density, then
$X, X^2, X^3, X^4, \ldots$ is Benford with probability 1. **(Berger-H).**

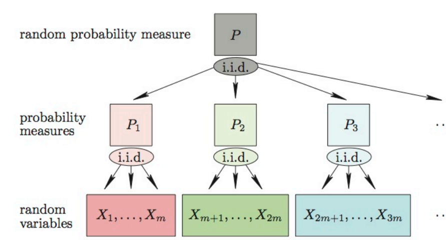
Thm 6. If $X_1, X_2, X_3, X_4, \ldots$ are i.i.d. random variables with a density, then
$X_1, X_1X_2, X_1X_2X_3, \ldots$ is Benford with probability 1. **(Berger-H).**



BL Probability Theorems (cont'd)

Mixing Data from Different Distributions

Thm 7. Combining random samples from unbiased random distributions yields a Benford distribution in the limit (with probability 1).



Ex.

Average.......	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error	±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Three Common Errors

1. ***Not all*** **exponential sequences $a, a^2, a^3, \ldots$ are Benford.**

   Ex. If $a = \sqrt{10}$,

   then the first digits of $a, a^2, a^3, \ldots$ are 3,1,3,1,3,1,...

2. ***No*** **sequence $a, 2a, 3a, 4a, \ldots$ (or sums of *iid* random variables) are Benford.**

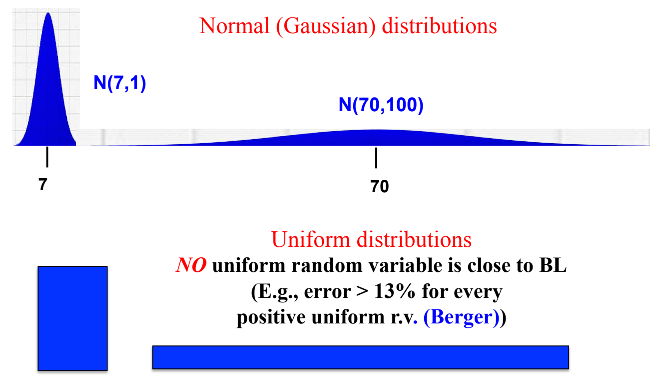
3. **A BL distribution need *not* cover many orders of magnitude.**

   Ex. If U is a Uniform(0,1) random variable, then

   $X = 10^U$ is **exactly** Benford,* and $1 \le X < 10$.

A *Widespread* Error

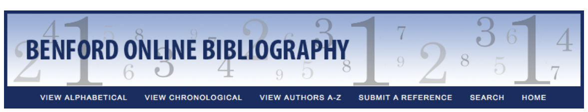
4. Regularity and large spread do *not* imply BL.



Online Resources

Free searchable Benford Online Bibliography:

http://www.benfordonline.net/



Open-access monograph: ***A basic theory of Benford's law*** **(Berger-H, 2011, Probability Surveys 8, 1-126)**

http://www.i-journals.org/ps/viewissue.php?id=11#Articles

Mathworld

http://mathworld.wolfram.com/BenfordsLaw.html

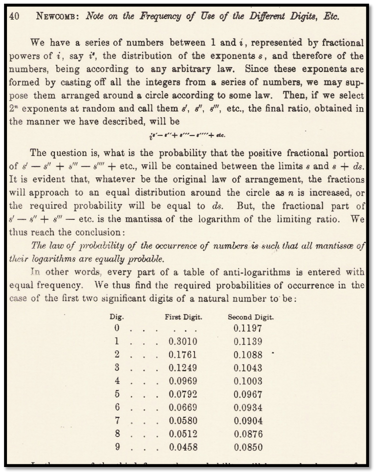
Thank you, European Commission!

And especially the organizers:

Domenico Perrotta, European Commission, Joint Research Centre, Italy (Chair)

Andrea Cerioli, Università di Parma, Italy

Lucio Barabesi, Università di Siena, Italy


40 NEWCOMB: *Note on the Frequency of Use of the Different Digits, Etc.*

We have a series of numbers between 1 and i, represented by fractional powers of i, say i^s, the distribution of the exponents s, and therefore of the numbers, being according to any arbitrary law. Since these exponents are formed by casting off all the integers from a series of numbers, we may suppose them arranged around a circle according to some law. Then, if we select 2^n exponents at random and call them s', s'', s''', etc., the final ratio, obtained in the manner we have described, will be

$$i^{s' - s'' + s''' - s'''' + \text{etc.}}$$

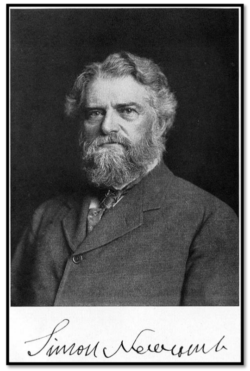
The question is, what is the probability that the positive fractional portion of $s' - s'' + s''' - s'''' +$ etc., will be contained between the limits s and $s + ds$. It is evident that, whatever be the original law of arrangement, the fractions will approach to an equal distribution around the circle as n is increased, or the required probability will be equal to ds. But, the fractional part of $s' - s'' + s''' -$ etc. is the mantissa of the logarithm of the limiting ratio. We thus reach the conclusion:

The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.

In other words, every part of a table of anti-logarithms is entered with equal frequency. We thus find the required probabilities of occurrence in the case of the first two significant digits of a natural number to be:

Dig.	First Digit.	Second Digit.
0	. . .	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Newcomb 1881



How to Win € from Friends

(Morrison, Ravikumar)

Players I and U each choose a positive integer.

Let *X* = product of the two integers.

I win if *X* begins with 1, 2, 3

U win if *X* begins with 4, 5, 6, 7, 8, or 9

We play 20 times –
winner gets €10 from loser each time.