WHAT?
WHEN?
WHY?

Q: What is Benford's Law?

A: It's all about quantities, GLORQ.

Q: When should data be Benford?

A: If order of magnitude is high and histogram falls to the right.

Q: Why does Benford's Law exist at all? Are there explanations?cuases?

A: There are (so far) 3 explanations for single issue data, such as earthquake data, population data, star mass data, expense data, income data, trade data, river flow data, etc. etc.

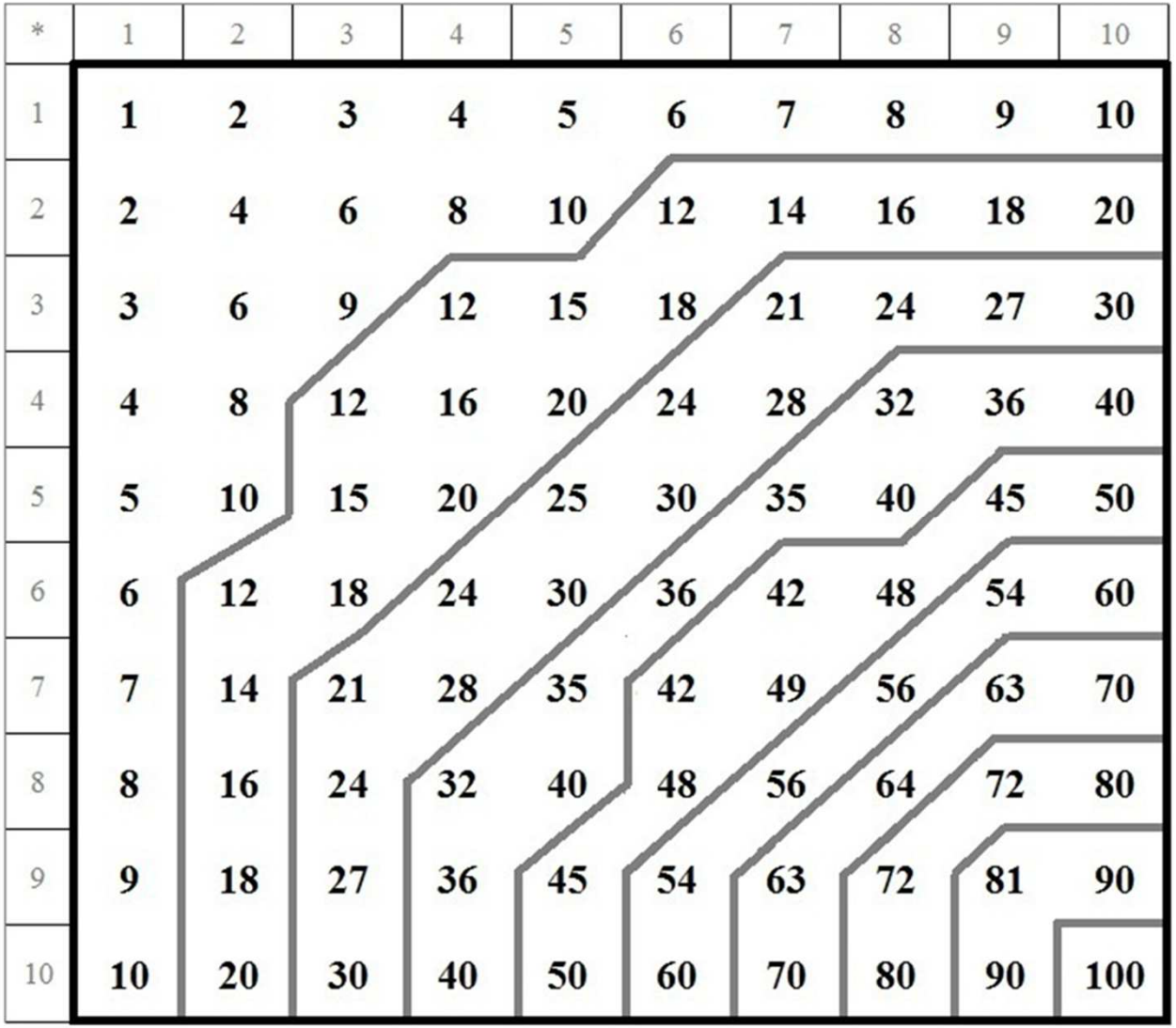
These 3 explanations are in addition and apart of the mixture of distributions explanation for the entire set of www data. Namely all existing numbers on the Internet, as well as all the numbers in all the books in all the libraries worldwide, and in all existing and past magazines and newspapers, etc. which are Benford as per Ted Hill explanation.

[i.e. that gigantic SETof ALL existing numbers on planet Earth.]

Multiplication

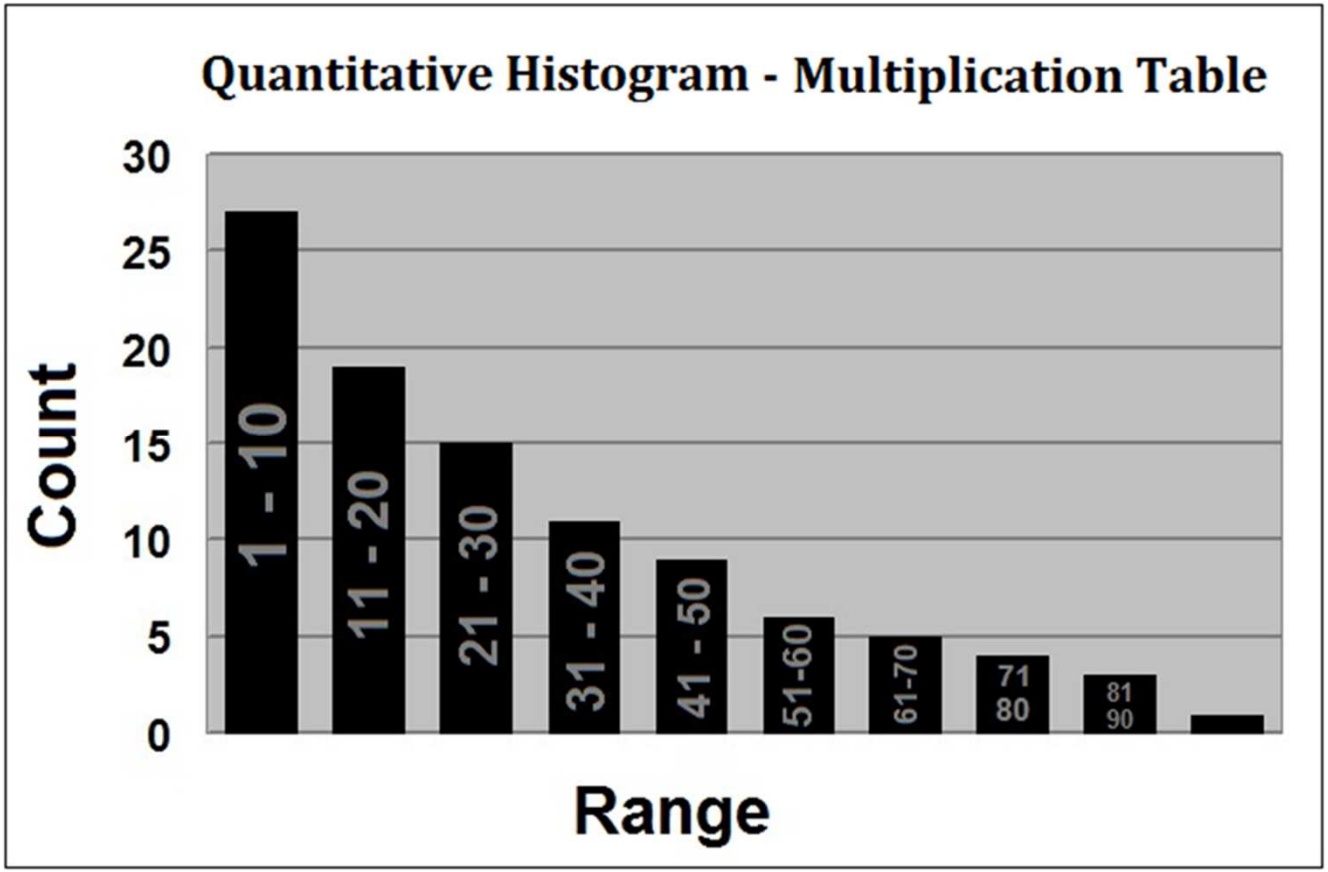
Multiplication processes produce sets of numbers favoring small quantities.

Small is Beautiful



*	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	10
2	2	4	6	8	10	12	14	16	18	20
3	3	6	9	12	15	18	21	24	27	30
4	4	8	12	16	20	24	28	32	36	40
5	5	10	15	20	25	30	35	40	45	50
6	6	12	18	24	30	36	42	48	54	60
7	7	14	21	28	35	42	49	56	63	70
8	8	16	24	32	40	48	56	64	72	80
9	9	18	27	36	45	54	63	72	81	90
10	10	20	30	40	50	60	70	80	90	100

Quantitative Territorial Partition of the 10 by 10 Table


Quantitative Histogram - Multiplication Table
Count
30
25
20
15
10
5
0
1 - 10
11 - 20
21 - 30
31 - 40
41 - 50
51-60
61-70
71
80
81
90
Range

Small is beautiful !

Central Limit Theorem

Adding many IID variables leads to the Normal distribution in the limit.

Sum = X_1+X_2+X_3+… = Normal(m, sd)

Multiplicative Central Limit Theorem

Multiplying many variables leads to the Lognormal distribution in the limit.

Product = X_1*X_2*X_3*… = Lognormal(s, l)

$$e^{X1+X2+X3+\ldots} = (e^{x1})(e^{x2})(e^{x3})\ldots$$

$$e^{Normal} = Lognormal$$

$$Lognormal = (e^{x1})(e^{x2})(e^{x3})\ldots$$

***Lognormal* with high shape parameter** *(shape > 1)* **is perfectly Benford for all practical purposes.**

Why?

Because the Lognormal is 'made of' multiplications !

On a more profound level, the typical multiplicative form of the equations in physics, chemistry, astronomy, and other disciplines, as well as those of their many applications and results, lead to the manifestation of Benford's Law in the physical world.

Newton gave us **$F = M*A$**,

not **$F = M + A$**.

He gave us **$F_G = G*M_1*M_2 / R^2$**,

not **$F_G = G + M_1 + M_2 - R^2$**

and such is the state of affair in so many other physical expressions.

Partitions

Partitioning as a Cause of the Small is Beautiful Phenomenon and Benford

‘One big quantity is composed of numerous small quantities’,

or equivalently:

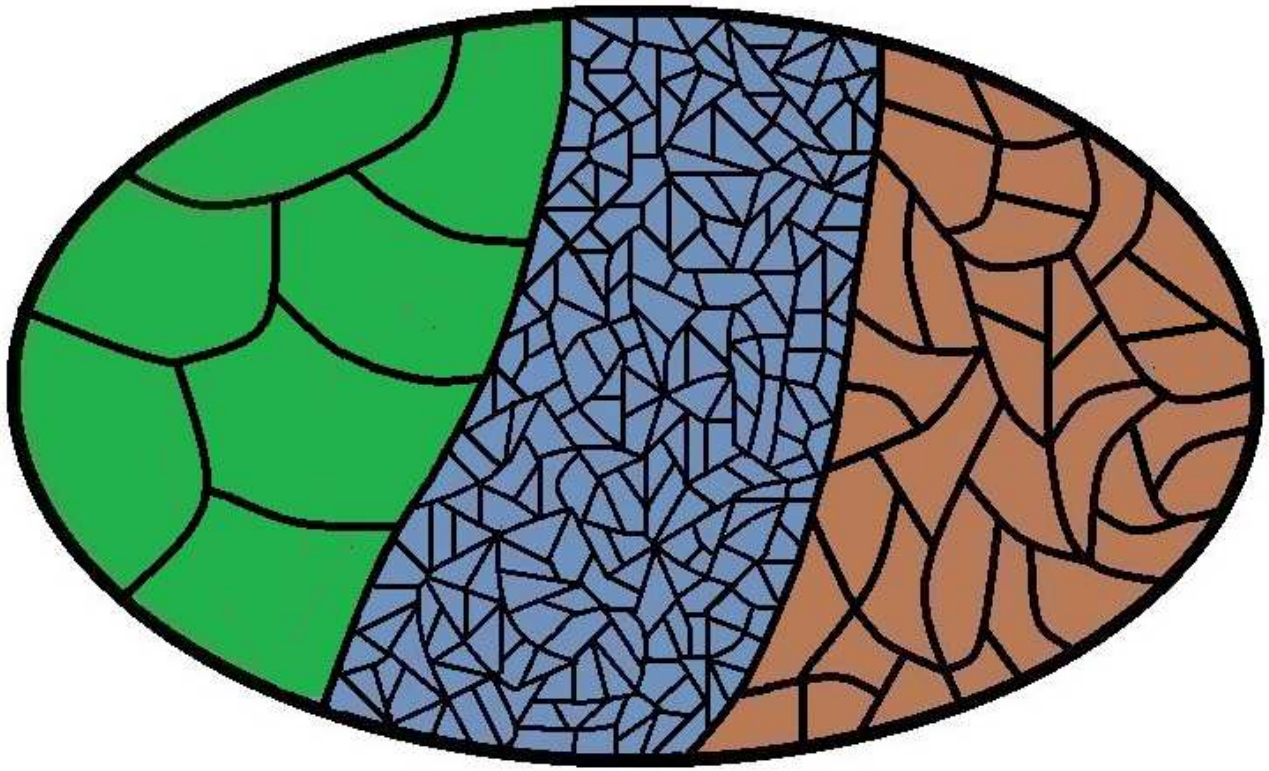
‘**Numerous small quantities are needed to merge into one big quantity**”.

The small is beautiful
in almost ALL partition models!

Benford's Law is found
in MANY partition models!

Italy is Partitioned into Either:

Few Big Parts **OR** Many Small Parts



An Equitable Mix of Small, Medium, and Big Yielding ‘Small is Beautiful’

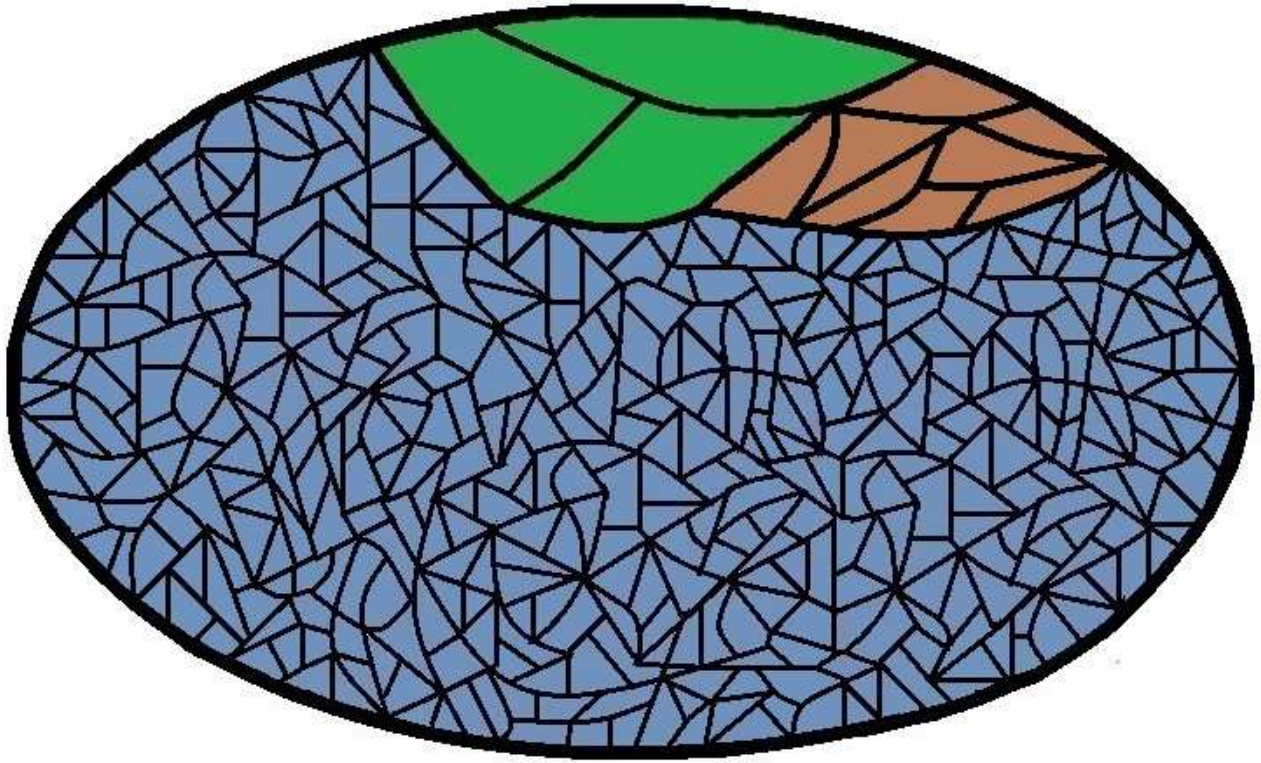
The previous figure depicts one possible random partition in the natural world where:

approximately **1/3** of the entire oval area consists of big parts (around the left side);

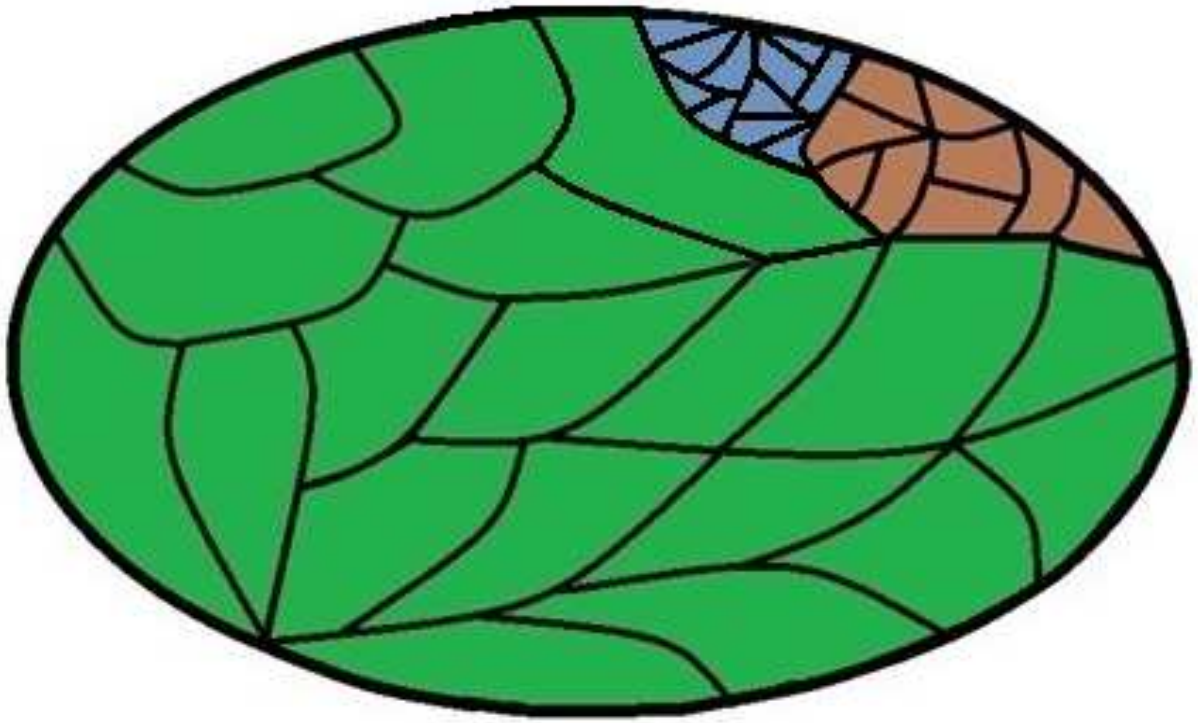
approximately **1/3** of the entire oval area consists of small parts (around the center);

approximately **1/3** of the entire oval area consists of medium parts (around the right side);

namely endowing equal portions of overall quantity **fairly** to each size **without any bias**.



Uneven Mix with Too Many Small Parts
Yielding ‘Small is extremely Beautiful’



Uneven Mix with Too Many Big Parts
Yielding Unnatural and Rare Configuration

Data Aggregation

Data Set A: {2, 3, 5, 7}
Data Set B: {1, 4, 6, 9, 13, 14}
Data Set C: {2, 6, 7, 9, 11, 15, 16, 21}
Data Set D: {1, 2, 6, 8, 13, 14, 19, 23, 25}
Data Set E: {3, 4, 8, 12, 15, 19, 22, 24, 29, 35, 41}
Data Set F: {1, 5, 8, 11, 12, 17, 19, 24, 27, 32, 38, 43, 47}

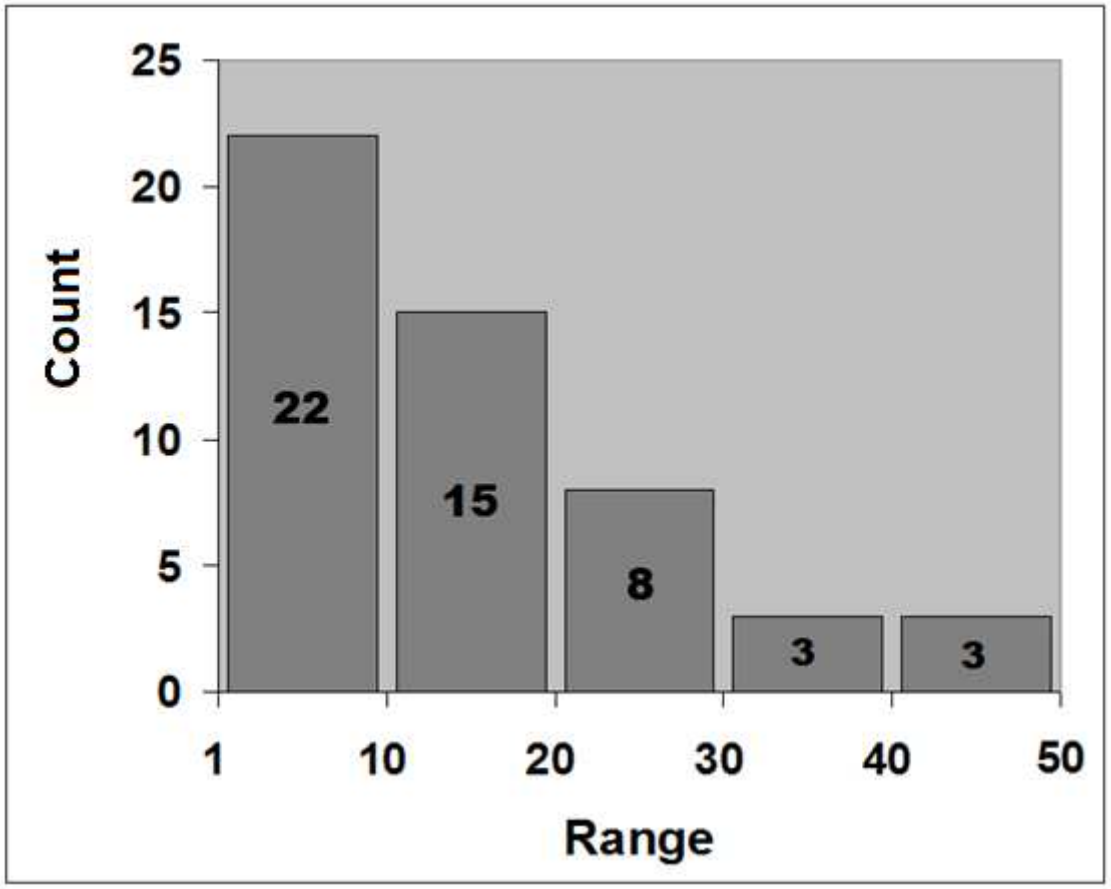
The combined data set A, B, C, D, E, F:

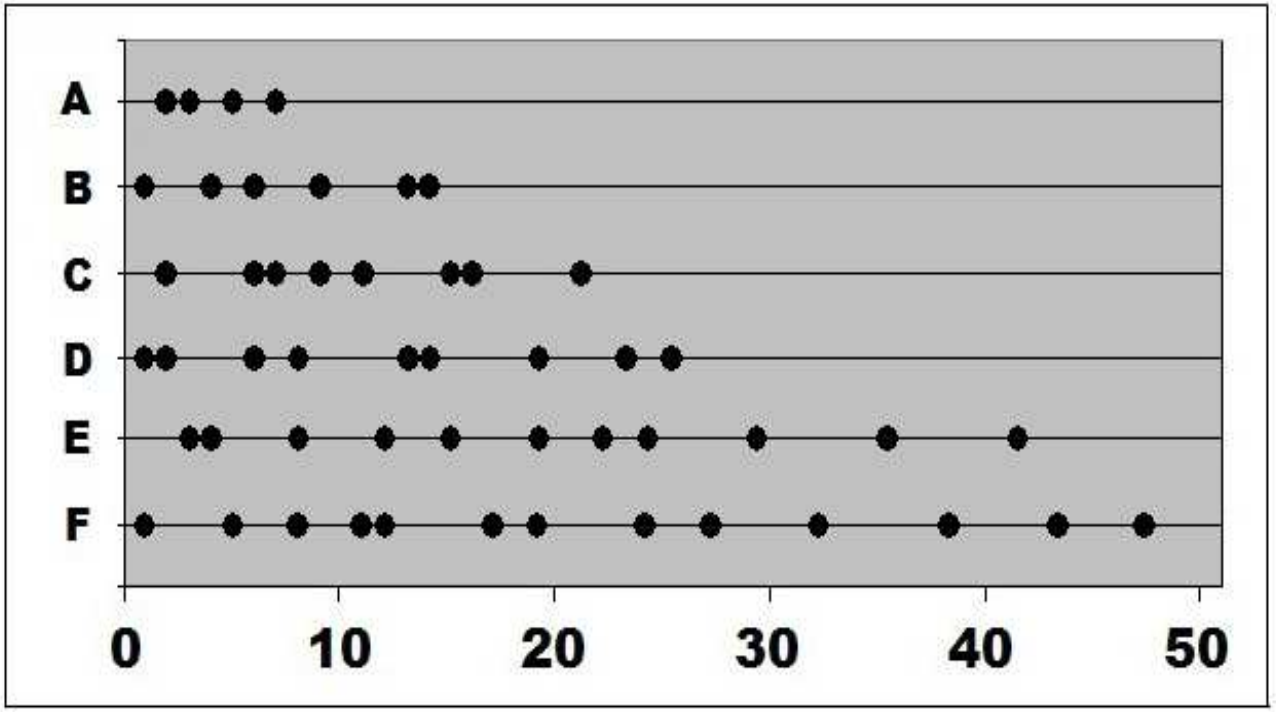
{2, 3, 5, 7, 1, 4, 6, 9, 13, 14, 2, 6, 7, 9, 11, 15, 16, 21, 1, 2, 6, 8, 13, 14, 19, 23, 25, 3, 4, 8, 12, 15, 19, 22, 24, 29, 35, 41, 1, 5, 8, 11, 12, 17, 19, 24, 27, 32, 38, 43, 47}

A, B, C, D, E, F, sorted, ordered from low to high:

{1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9,
11, 11, 12, 12, 13, 13, 14, 14, 15, 15, 16, 17, 19, 19, 19,
21, 22, 23, 24, 24, 25, 27, 29, 32, 35, 38, 41, 43, 47}

Histogram:





Small is beautiful!

END

chi-sqt Test is not appropriate for Benford's Law!

compliance vs. comparison

"**For this sample drawn from a supposedly logarithmic population, is digital deviation from the logarithmic due to chance or structural?**"

"**How far from the logarithmic is this digital configuration?**"

Almost in all cases, the underlying theoretical and statistical basis for the chi-sqr test are **not applicable** to the data set under consideration, as seen by its supposed **“oversensitivity”** in the cases of large data sets where even mild deviations from the logarithmic are flagged as significant!

Yet, it has erroneously been used in accounting and auditing circles on a regular basis for many years, and unfortunately it is still being used nowadays as part of the standard procedure in fraud detection.

This has led to much confusion and many errors, and has done a lot in general to undermine trust in the whole discipline of Benford's Law.

Even more unfortunate is its use in mathematical and empirical research where it is also erroneously applied blindly in almost all cases, lacking statistical justification, **and has lead to numerous misguided conclusions and much confusion.**

Instead of the chi-sqr

We use subjective yet reasonble guidelines

Sum of Squares Deviation Measure (SSD)

$$SSD = \sum(\text{Observed \%} - \mathbf{100}*LOG(1+1/d))^{\mathbf{2}}$$

For example:

$$\mathbf{SSD} = (31.1 - \mathbf{30.1})^2 + (18.2 - \mathbf{17.6})^2 + (13.3 - \mathbf{12.5})^2 +$$
$$(9.4 - \mathbf{9.7})^2 + (7.2 - \mathbf{7.9})^2 + (6.3 - \mathbf{6.7})^2 +$$
$$(5.9 - \mathbf{5.8})^2 + (4.5 - \mathbf{5.1})^2 + (4.1 - \mathbf{4.6})^2$$
$$= \mathbf{3.4}$$

Empirical Rule of Compliance

SSD generally should be below **25**;

A data set with SSD over **100** is considered to deviate too much from Benford;

And a reading below **2** is considered to be ideally Benford.