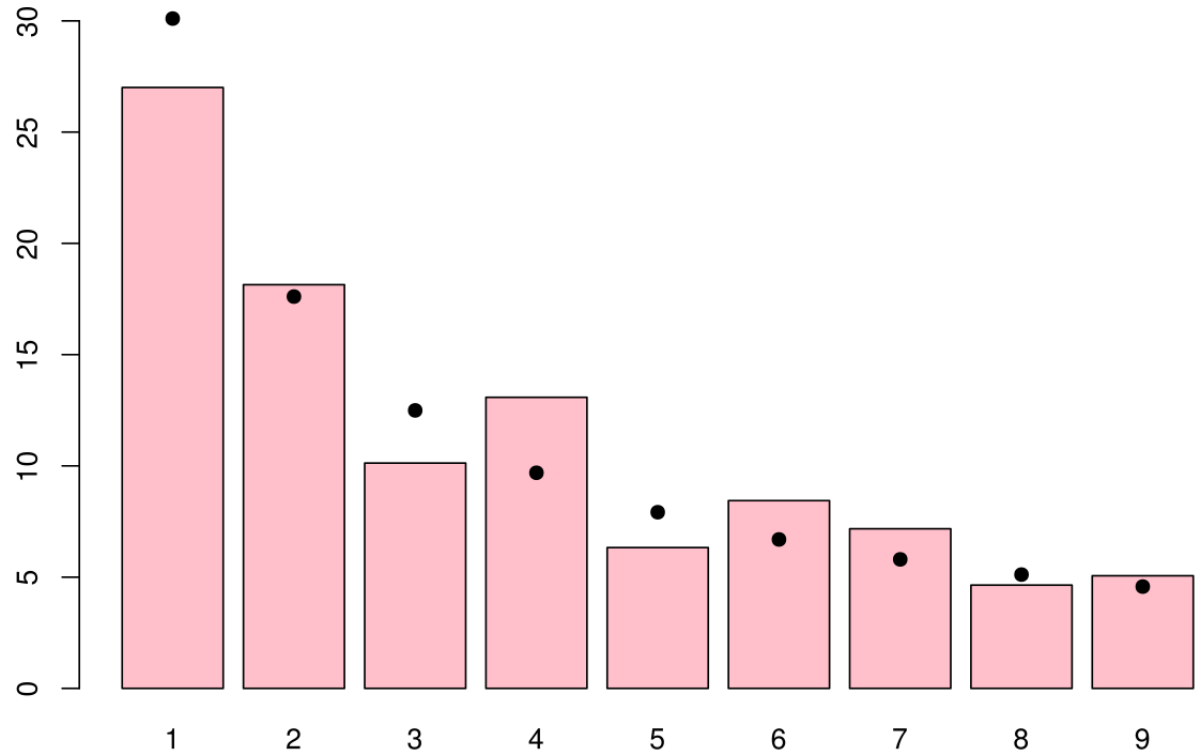
Mathematical Laws in Data Science

Asim Tewari
Professor Department of Mechanical Engineering
IIT Bombay, Powai, Mumbai 400 076,
India

Population of countries



Distribution of first digits (in %, red bars) in the population of the 237 countries of the world as of July 2010. Black dots indicate the distribution predicted by Benford's law.

Heights of tallest structures

Leading digit	m		ft		Per Benford's law
	Count	%	Count	%	
1	24	41.4 %	16	27.6 %	30.1 %
2	9	15.5 %	8	13.8 %	17.6 %
3	7	12.1 %	5	8.6 %	12.5 %
4	6	10.3 %	7	12.1 %	9.7 %
5	1	1.7 %	10	17.2 %	7.9 %
6	5	8.6 %	4	6.9 %	6.7 %
7	1	1.7 %	2	3.4 %	5.8 %
8	4	6.9 %	5	8.6 %	5.1 %
9	1	1.7 %	1	1.7 %	4.6 %

Examining a list of the heights of the 58 tallest structures in the world by category shows that 1 is by far the most common leading digit, irrespective of the unit of measurement

Leading digit of 2^n

Leading digit	Occurrence		Per Benford's law
	Count	%	
1	29	30.2 %	30.1 %
2	17	17.7 %	17.6 %
3	12	12.5 %	12.5 %
4	10	10.4 %	9.7 %
5	7	7.3 %	7.9 %
6	6	6.3 %	6.7 %
7	5	5.2 %	5.8 %
8	5	5.2 %	5.1 %
9	5	5.2 %	4.6 %

Another example is the leading digit of 2^n. The sequence of the first 96 leading digits (1, 2, 4, 8, 1, 3, 6, 1, 2, 5, 1, 2, 4, 8, 1, 3, 6, 1,..)

Benford's Law

- Benford's law, also called the **Newcomb–Benford law**, the **law of anomalous numbers**, or the **first-digit law,** is a mathematical law about the leading digit number in a real-world dataset.
- Benford law state that a set of numbers is said to satisfy Benford's law if the leading digit *d* ($d \in 1, \ldots,$

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1 + \frac{1}{d}\right)$$

Benford's Law

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right)$$

The leading digits in such a set thus have the following distribution:

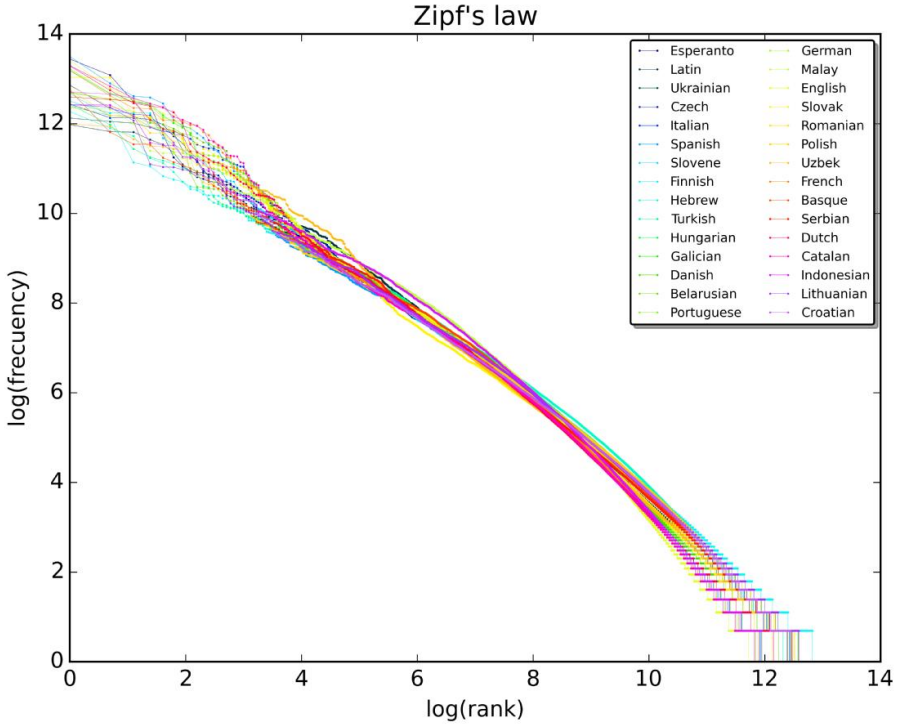
d	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

Zipf's Law

- Zipf’s law was created for quantitative linguistics and states that given some natural language dataset corpus, any word's frequency **is inversely proportional to its frequency table rank**. Thus the most frequent word occurs approximately twice as often as the second most frequent word and three times as often as the third most frequent word.

Zipf's Law

- Zipf’s law was created for quantitative linguistics and states that given some natural language dataset corpus, any word's frequency **is inversely proportional to its frequency table rank**. Thus the most frequent word occurs approximately twice as often as the second most frequent word and three times as often as the third most frequent word.



A plot of the rank versus frequency for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a log-log scale.