自然语言理解与机器翻译

第二章：自然语言的统计特性

刘均（liukeen@xjtu.edu.cn）

陕西省天地网技术重点实验室

智能网络与网络安全教育部重点实验室

1 2.1 Zipf定律

2 2.2 Heaps定律

3 2.3 Benford定律

基本要求：掌握Zipf定律、Heaps定律、Benford定律及其在NLU中的应用。

2.1 Zipf定律

词频分布

- ✓ 美国语言学家 George Kingsley Zipf
- ✓ 少数词非常普遍：
  - 最常见的250~300个英文单词超过了文本中单词数量的50%
  - the与 of 占了 10%， and, to, a, in占了 10%，接下来的12个单词占了 10%.
  - 《白鲸》（Moby Dick）小说的第一章，词汇表大小是859 unique words (types), 单词数是2256 occurrences

http://www.cs.yale.edu/homes/radev/nlpclass/slides2018/144.pptx

2.1 Zipf定律

词频分布

- ✓ 大多数出现的频次非常低：一半左右的词在语料库中只出现一次
- ✓ 词频是长尾（long tailed）或重尾（heavy tailed）分布

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

2.1 Zipf定律

- ✓ Rank (r): 一个词按照词频(f) 从大到小的排列次序
- ✓ George Kingsley Zipf (1902–1950) 发现:

$$f \cdot r = c \quad (\ for\ constant\ c)$$

- The *i* th most frequent term has frequency proportional to *1/i*
- Let this frequency be *c/i.*
- Then $\sum_{i=1}^{500,000} c/i = 1.$
- The *k* th Harmonic number is $H_k = \sum_{i=1}^{k} 1/i.$
- $c = 1/H_m = 1/\ln(500k) \sim 1/13.$
- So the *i* th most frequent term has frequency roughly *1/13i.*

2.1 Zipf定律

the	1130021	from	96900	or	54958
of	547311	he	94585	about	53713
to	516635	million	93515	market	52110
a	464736	year	90104	they	51359
in	390819	its	86774	this	50933
and	387703	be	85588	would	50828
that	204351	was	83398	you	49281
for	199340	company	83070	which	48273
is	152483	an	76974	bank	47940
said	148302	has	74405	stock	47401
it	134323	are	74097	trade	47310
on	121173	have	73132	his	47116
by	118863	but	71887	more	46244
as	109135	will	71494	who	42142
at	101779	say	66807	one	41635
mr	101679	new	64456	their	40910
with	101210	share	63925		

Frequency of 50 most common words in English (sample of 19 million words)

2.1 Zipf定律

rf*1000/n

the	59	from	92	or	101
of	58	he	95	about	102
to	82	million	98	market	101
a	98	year	100	they	103
in	103	its	100	this	105
and	122	be	104	would	107
that	75	was	105	you	106
for	84	company	109	which	107
is	72	an	105	bank	109
said	78	has	106	stock	110
it	78	are	109	trade	112
on	77	have	112	his	114
by	81	but	114	more	114
as	80	will	117	who	106
at	80	say	113	one	107
mr	86	new	112	their	108
with	91	share	114		

2.1 Zipf定律

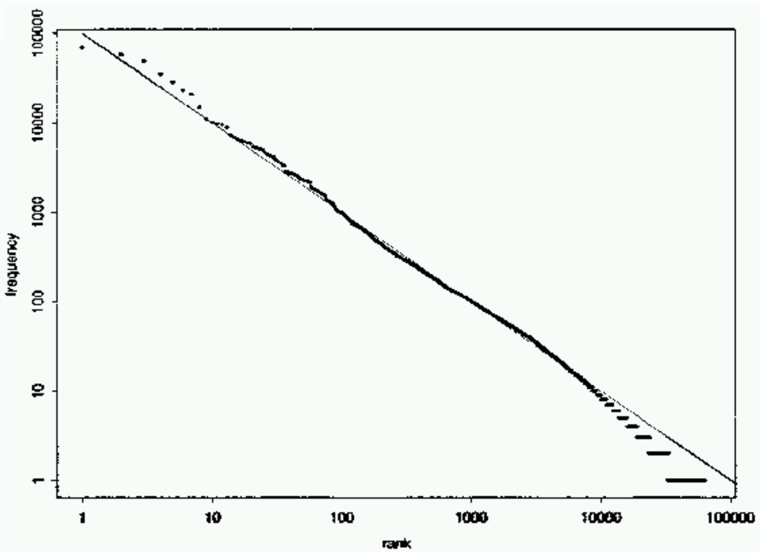
Zipf定律的符合程度

- ✓ 符合$y = kx^c$的分布为幂律分布 (power law)
- ✓ 在双对数坐标系(log–log plot)，幂律分布为斜率为 *c*的直线

$$\log(y) = \log(kx^c) = \log k + c\log(x)$$

- ✓ 除了rank很高与很低的术语，Zipf 分布与实际符合较好

Fit to Zipf for Brown Corpus

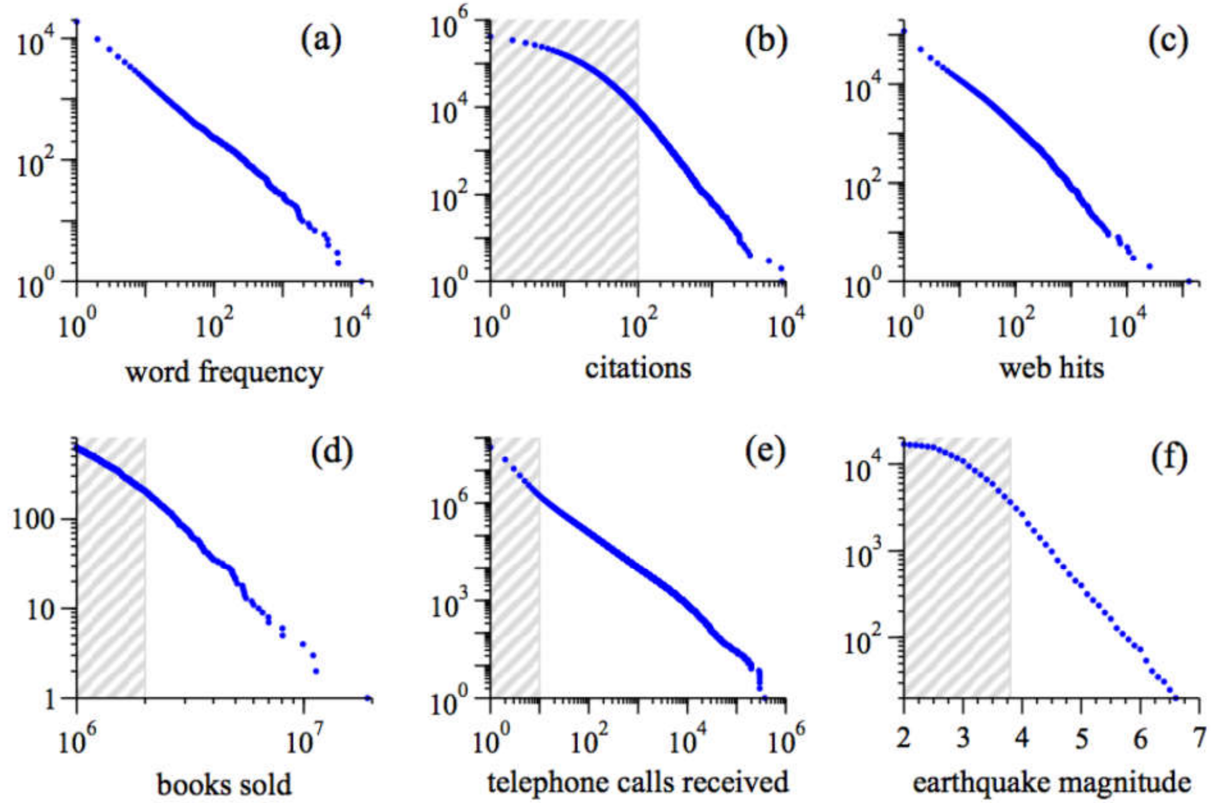


2.1 Zipf定律

Zipf定律实验

2.1 Zipf定律

■ **幂律分布很常见**



https://statweb.stanford.edu/~owen/courses/306a/ZipfByHera.pdf

2.1 Zipf定律

Zipf定律的解释

✓ **Miller's Monkey**

- 长度为 *i* 的词的概率： $P(i) = (1/27)^i (1/27) = (1/27)^{i+1}$
- 长度为 *i* 的词的*rank*值 $\sum_{j=1}^{i-1} 26^j < r_i \leq \sum_{j=1}^{i} 26^j$

$$r = \sum_{j=1}^{i} 26^j = \frac{26}{25}(26^i - 1),$$

$$i' = \frac{\log\left(\frac{25}{26}r + 1\right)}{\log 26}$$

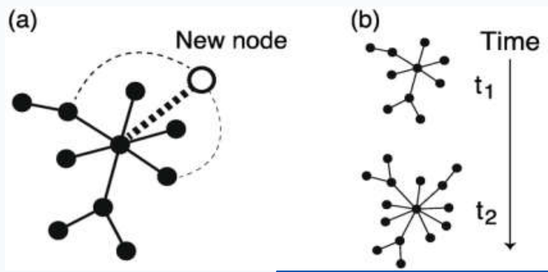
$$\begin{aligned} p(i') &= (1/27)^{i'+1} \\ &= (1/27)^{\frac{\log\left(\frac{25}{26}r+1\right)}{\log 26}+1} \\ &= (1/27)\left(\frac{25}{26}r + 1\right)^{-\frac{\log 27}{\log 26}} \quad \text{using the fact } a^{\log b} = b^{\log a} \\ &\approx 0.04(r + 1.04)^{-1.01}, \end{aligned}$$

参考http://pages.cs.wisc.edu/~jerryzhu/cs838/words.pdf

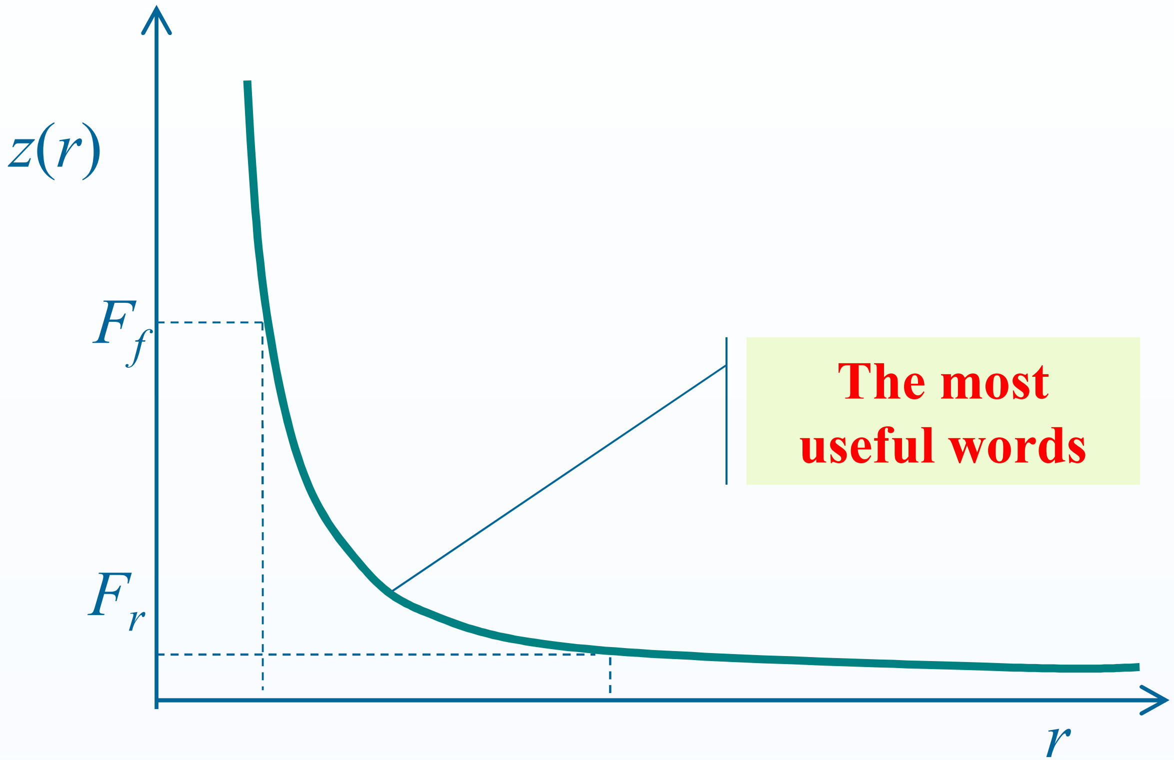
2.1 Zipf定律

■ Zipf定律的解释

- ✓ **Principle of least effort**：词频的差异有助于使用较少的词汇表达尽可能多的语义
- ✓ **语言使用的影响机制**：Preferential attachment
  - Start with a limited number of initial nodes m_0
  - At each time step, add a new node that has m edges that link to m existing nodes in the system $m \le m_0$
  - When choosing the nodes to which to attach, assume a probability ∏ for a node i proportional to the number *ki* of links already attached to it $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$
  - After t time steps, the network will have $n=M+m_0$ nodes and $M=mt$ edges $n = t + m_0$, $M = mt$
  - It can be shown that this leads to a power law network!



2.1 Zipf定律



Luhn (1958) suggested that both extremely common and extremely uncommon words were not very useful for indexing.

2.1 Zipf定律

- **Zipf定理对索引的作用**
  - ✓ 好的索引词汇
    - 词频太高：可能返回所有文档
    - 词频太低： 仅能返回很少的文档
  - ✓ 利用Zipf’s Law可以去除频次高的Stopword，优化的倒排索引的时空开销

2.1 Zipf定律

Doc 1
one fish, two fish

Doc 2
red fish, blue fish

Doc 3
cat in the hat

Doc 4
green eggs and ham

	1	2	3	4
blue		1		
cat			1	
egg				1
fish	1	1		
green				1
ham				1
hat			1	
one	1			
red		1		
two	1			

blue	→ 2	
cat	→ 3	
egg	→ 4	
fish	→ 1	→ 2
green	→ 4	
ham	→ 4	
hat	→ 3	
one	→ 1	
red	→ 2	
two	→ 1	

倒排索引

2.2 Heaps定律

- Heaps在1978年提出。他观察到在语言系统中，词汇表的大小与文本篇幅（所有出现的单词累积数目）之间存在幂函数关系，其幂指数小于1
- Heaps 定理的作用
  - ✓ 可以估测跟定文本集的词汇表的大小
  - ✓ 可以预测随着文本集增长倒排索引规模的变化

2.2 Heaps定律

- **Heaps定律：**

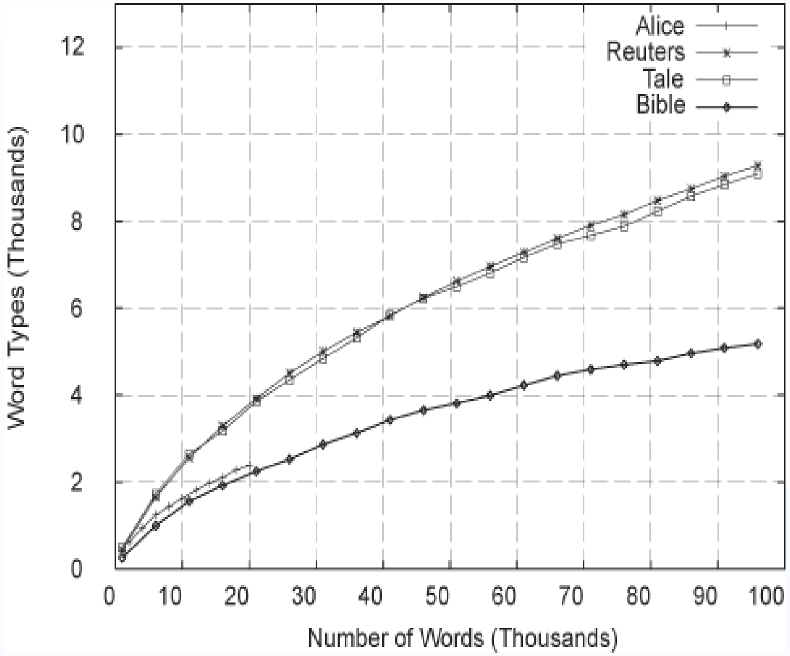
$$V = Kn^{\beta} \quad \text{with constants } K,\ 0<\beta<1$$

✓ V是词汇表的大小

✓ n是文本集词的个数

✓ $10 \le K \le 100$, $\beta \approx 0.4 \sim 0.6$

✓ 目前最匹配的$K=44$, $\beta=0.49$



✓ 在Reuters-RCV1 前1,000,020 词的数据集中预测词汇大小为 38,323，实际为38,365

Heaps定律实验

2.2 Heaps定律

- We want to estimate the size of the vocabulary for a corpus of 1,000,000 words. However, we only know statistics computed on smaller corpora sizes:
  - ✓ For 100,000 words, there are 50,000 unique words
  - ✓ For 500,000 words, there are 150,000 unique words
  - ✓ Estimate the vocabulary size for the 1,000,000 words corpus
  - ✓ How about for a corpus of 1,000,000,000 words?

2.3 Benford定律 (第一数字定律)

- 现实世界的数据集中, 首位数字为1至9的样本数量并非均匀分布，而是从1至9随着数值增加，频率逐渐减少
- 1881年，天文学家 **Simon Newcomb**发现对数表中以1起首的页较为破旧。
- 1938年，物理学家 **Frank Benford**做了**20,229** 组观测（人口数量、出生率、物理化学常数等），给出**Benford定律**
- 1938年，WVU的**Mark Nigrini**将Benford定律作为审计工具，检测公司数据的异常

2.3 Benford定律

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

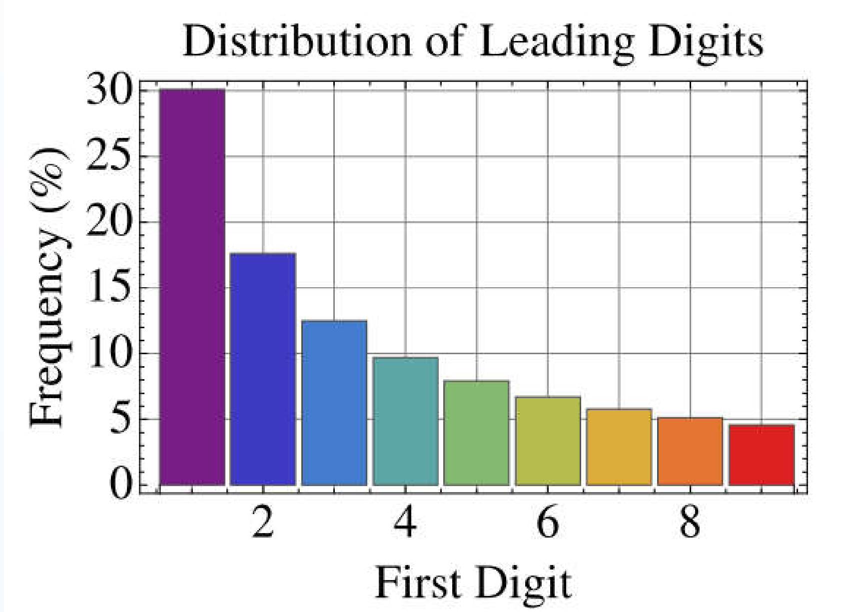
首位数字	比例
1	30.1%
2	17.6%
3	17.6%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

适用性：

① d 大于9时仍适用；

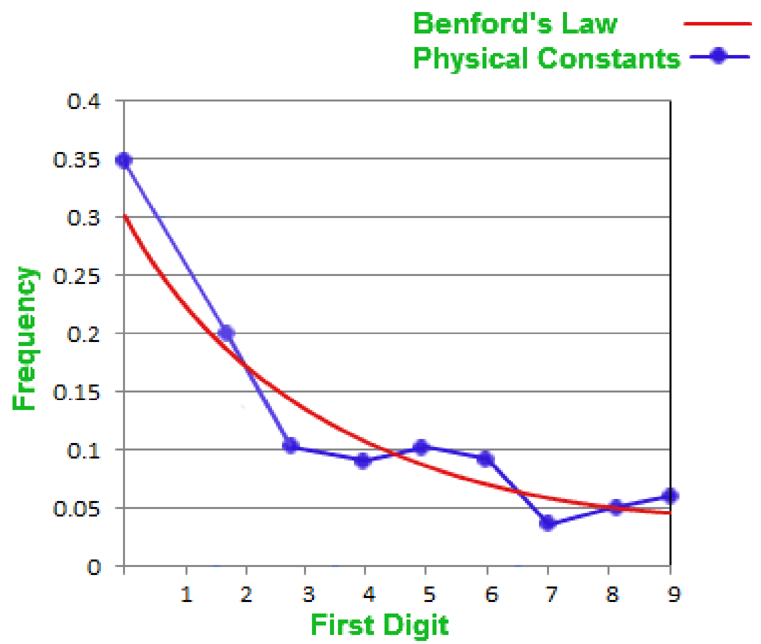
② 数据不是十进制仍适用；

③ 数量级跨度越大越符合，整个国家的家庭收入/一个村的家庭收入

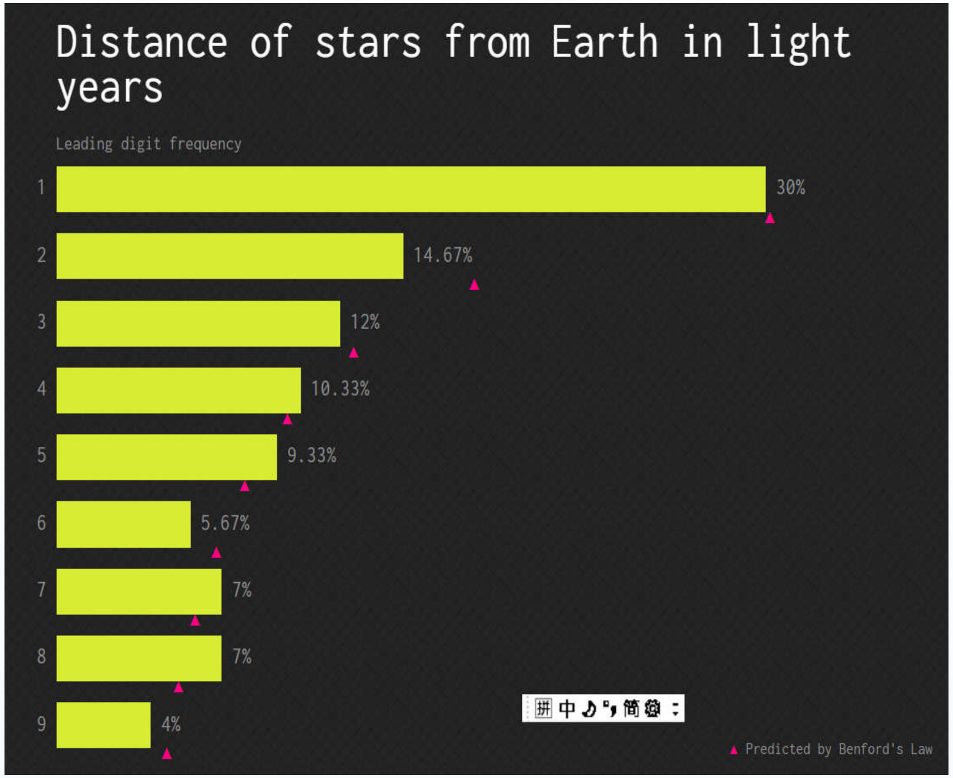


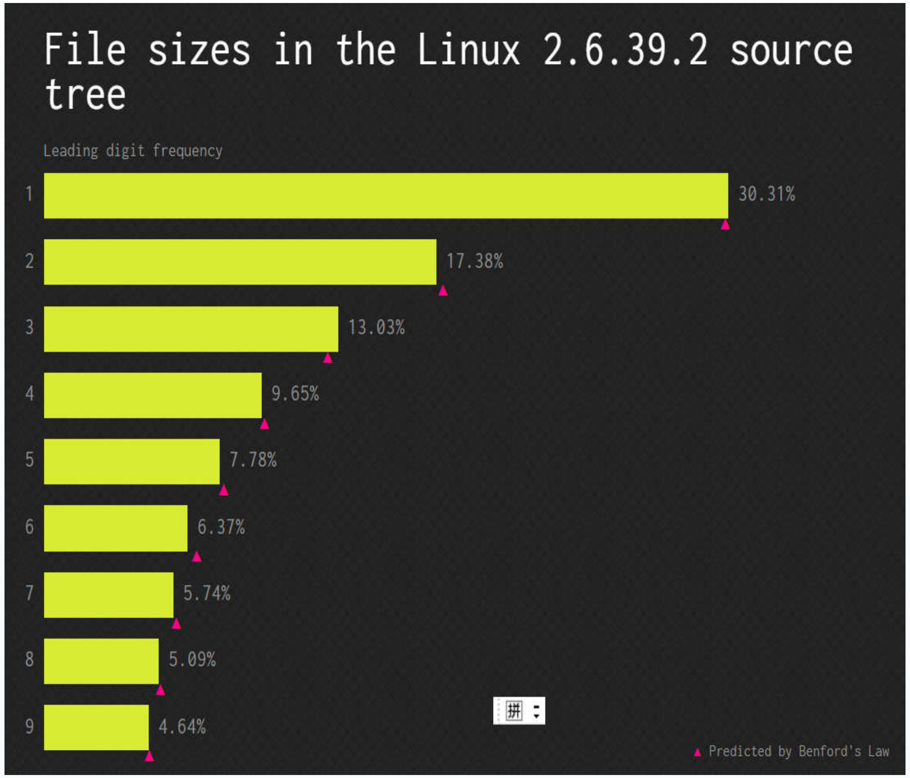
2.3 Benford定律 – 适用范围

- 由度量单位制获得的数据：人口数量、股票价格、半衰期、物理书中的答案、素数、Fibonacci数列、任何数字的幂（如2、3）
- 不符合数据主要是任意获得的和受限数据，如彩票数字、电话号码、汽油价格、日期、体重、身高



2.3 Benford定律





http://www.testingbenfordslaw.com/google-books-1-grams

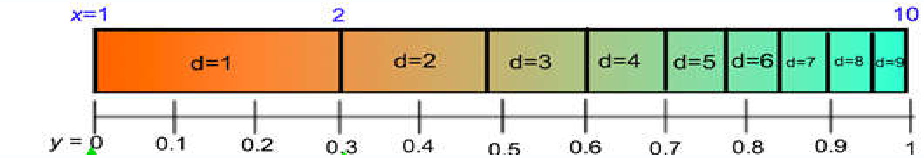
2.3 Benford定律 – 两个小实验

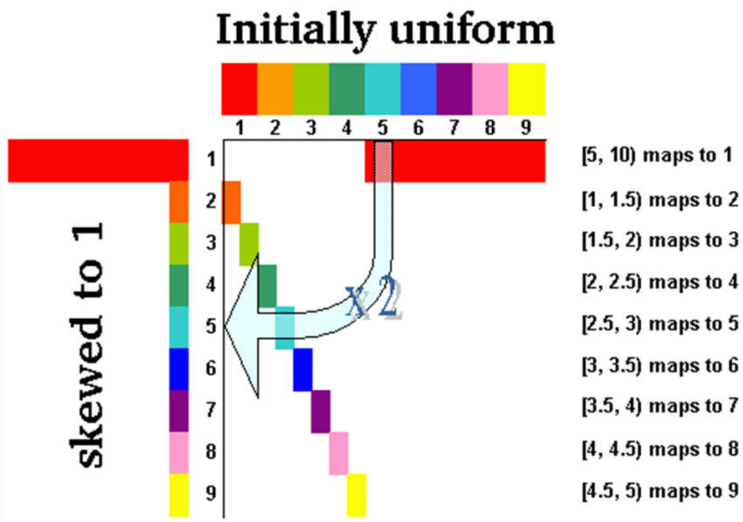
- 实验1： 文本集中的所有数字
- 实验2： Fibonacci 数列

$$F_n = \begin{cases} 0 & \text{if } n = 0; \\ 1 & \text{if } n = 1; \\ F_{n-1} + F_{n-2} & \text{if } n > 1. \end{cases}$$

2.3 Benford定律 －可能成因

- 固定倍率增长的数据，由数字a增长到a + 1起首的数的时间比a + 1到a + 2需要更多时间。（CPU主频、证券市场指数）





- 尺度不变形

- 对Benford定律的解释仍是开放问题。**不影响其应用。**

2.3 Benford定律 - 应用

Table 1. One of the columns gives the land area of political states and territories in km^2. The other column contains faked data, generated with a random number generator.

State/Territory	Real or Faked Area (km^2)	
Afghanistan	645,807	796,467
Albania	28,748	9,943
Algeria	2,381,741	3,168,262
American Samoa	197	301
Andorra	464	577
Anguilla	96	82
Antigua and Barbuda	442	949
Argentina	2,777,409	4,021,545
Armenia	29,743	54,159
Aruba	193	367
Australia	7,682,557	6,563,132
Austria	83,858	64,154
Azerbaijan	86,530	71,661
Bahamas	13,962	9,125
Bahrain	694	755
Bangladesh	142,615	347,722
Barbados	431	818
Belgium	30,518	47,123
Belize	22,965	20,648
Benin	112,620	97,768
...	...	...

https://statweb.stanford.edu/~owen/courses/306a/ZipfByHera.pdf