幂律与富者更富

及其与长尾、齐普夫定律等的关系

（基于第18章）

Power Law, Scale Free, Rich gets richer

Long Tail, Zipf’s Law，2/8 Law

流行性（popularity）

- 同一类事物的不同实例被关注、认知、或偏爱的程度
  - 人（明星），书籍，歌曲，某一类产品（例如软饮料），某一类服务（例如提供同一种服务的网站），微博主
- 为什么会有差别？
- 这种差别有没有什么规律？
- 有没有办法增进某些实例在这种差别中的优势？

流行性的定量观察

- 给定一个国家（地区）的网页集合（S），其中一个网页的入向链接数为 k 的概率 f(k) 是多少？
- 考虑在卓越和当当上销售的书籍集合（S），在其中发现销量为 k 的书的概率 f(k) 是多少？

– 它们的概率函数是否有相似之处，是否反映了一种规律，普适于其他具有流行现象的事物？
– 如果体现了反映流行现象的一种规律，为什么会有这规律？

以回答第一个问题为例

- 给定一个国家（地区）的网页集合（S），发现其中一个网页的入向链接数为 k 的概率 f(k) 是多少？

$$S=\{x_1^{(p_1)},x_2^{(p_2)},\ldots x_i^{(p_i)},\ldots,x_n^{(p_n)}\}$$

n是网页总数
p_i表示x_i的入向链接数

$$f(k)=\frac{\sum_{i=1}^{n} equal(p_i,k)}{n}$$

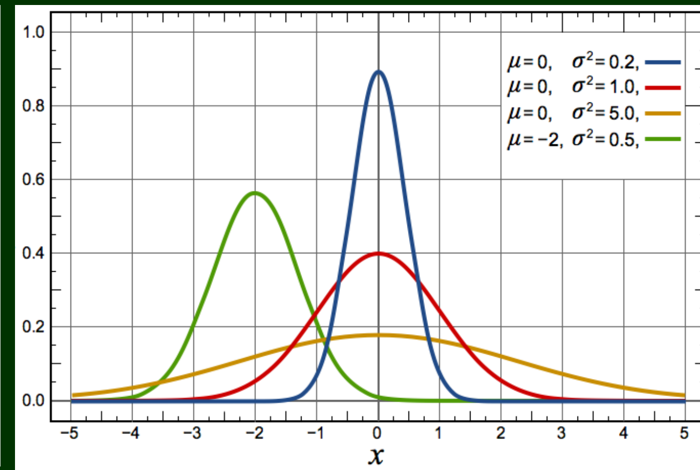
什么性质？
曲线是什么形状？

为什么不是正态分布？

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

概率密度函数

μ: 均值；σ²: 方差；σ: 标准差



中心极限定理：大量独立同分布的随机变量之和（均值）是正态分布的随机变量；与原始分布是什么无关。

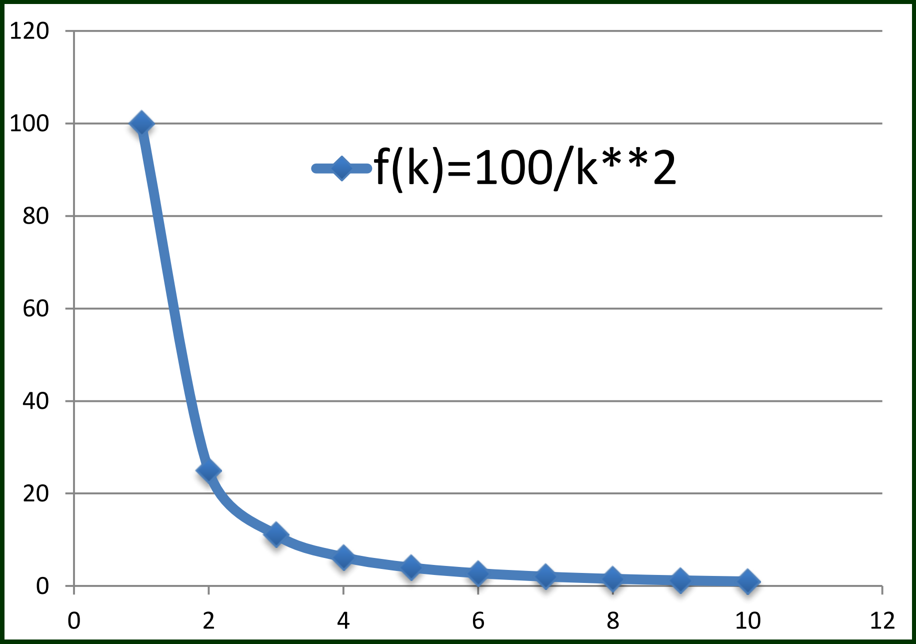
网页入向链接的个数（随机量）应该是什么分布？如果想象：网页A是否给网页B链接是一个随机变量；那么，B得到的入链个数就是大量随机变量之和。于是，正态分布？

数据实验表明：

$$f(k)=\frac{a}{k^c}=a\times k^{-c}$$

- 大量各种不同的数据集都显现出这种性态
- 因此，我们说这就是反映网页入度分布的规律，由于是幂函数，俗称“幂律”

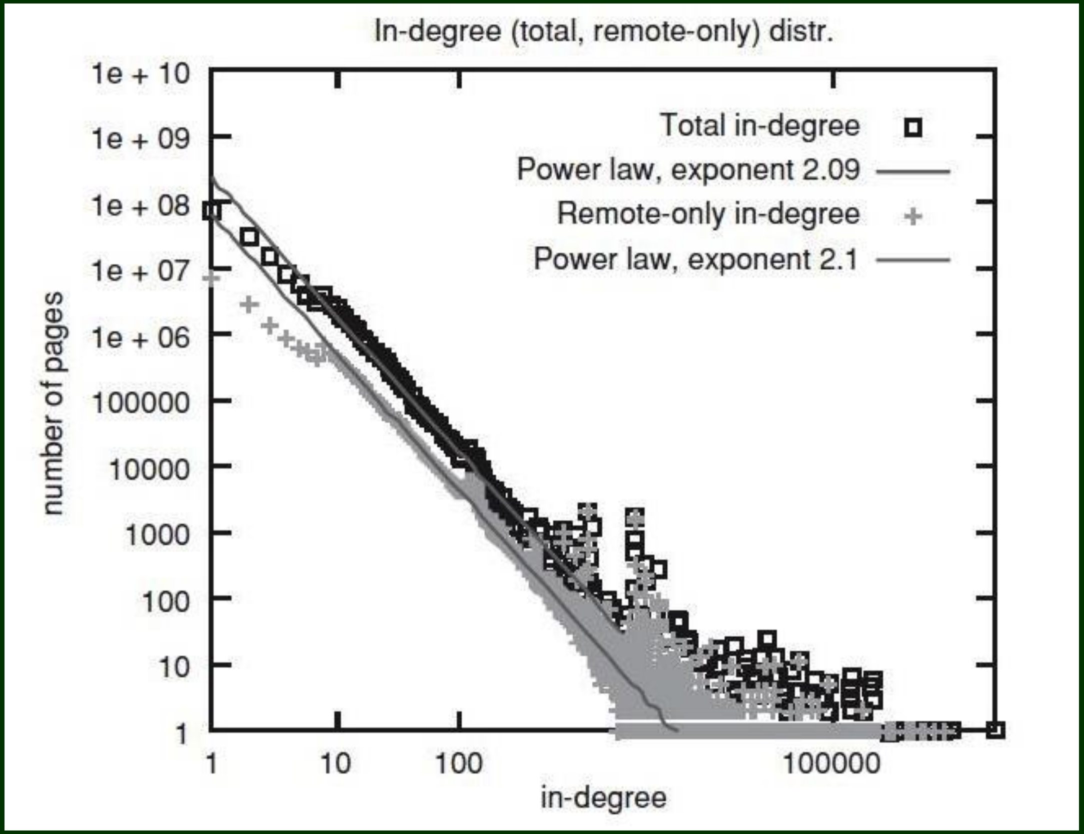
k	f(k)=1/k**2	g(k)=1/2**k
1	1	0.5
2	0.25	0.25
3	0.111111111	0.125
4	0.0625	0.0625
5	0.04	0.03125
6	0.027777778	0.015625
7	0.020408163	0.0078125
8	0.015625	0.00390625
9	0.012345679	0.001953125
10	0.01	0.000976563



幂率的习惯（图形）表示

$$f(k)=\frac{a}{k^{c}}=a\times k^{-c};\qquad \log(f(k))=\log(a)-c\times\log(k)$$

- log(f(k)) 是关于 log(k) 的线性函数
  - 以 log(k) 为横轴，log(f(k)) 为纵轴的图像是一条直线
- 这等价于说
  - 在对数坐标（横和纵）下，函数的图像是一条直线



因此，给定一组原始数据

k: 1, 2, 3, …
f(k): …

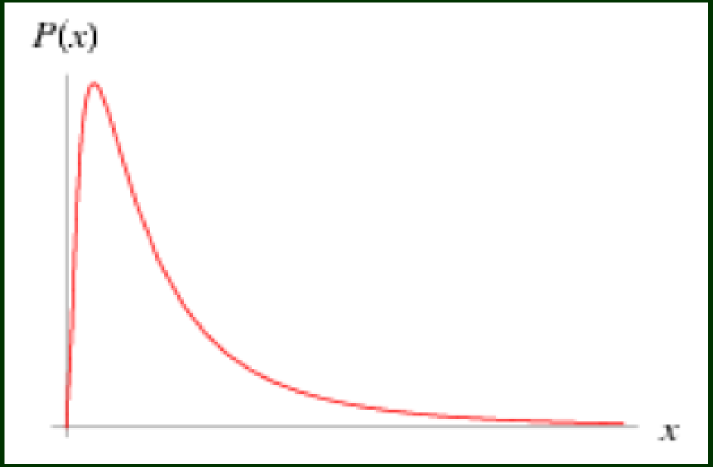
- 为查看f(k)是否幂律，一种做法就是取 log(k) 和对应的 log(f(k))，然后用得到的数据值在常规坐标下绘制曲线图形，观察结果看起来像不像一条直线。
- 在数据量很大的时候（流行度数据常常如此），这种方式很有效。许多绘图工具直接支持对数坐标。

幂律：流行度的一种主导规律

- 网页（网站）的入度，网站的出度
- 网站的规模（其中网页的数量）
- 每天能接到k个电话的电话
- 书籍的销量
- …

但不是100%普适的规律。
对数正态分布（log normal）也反映某些事物流行的现象。

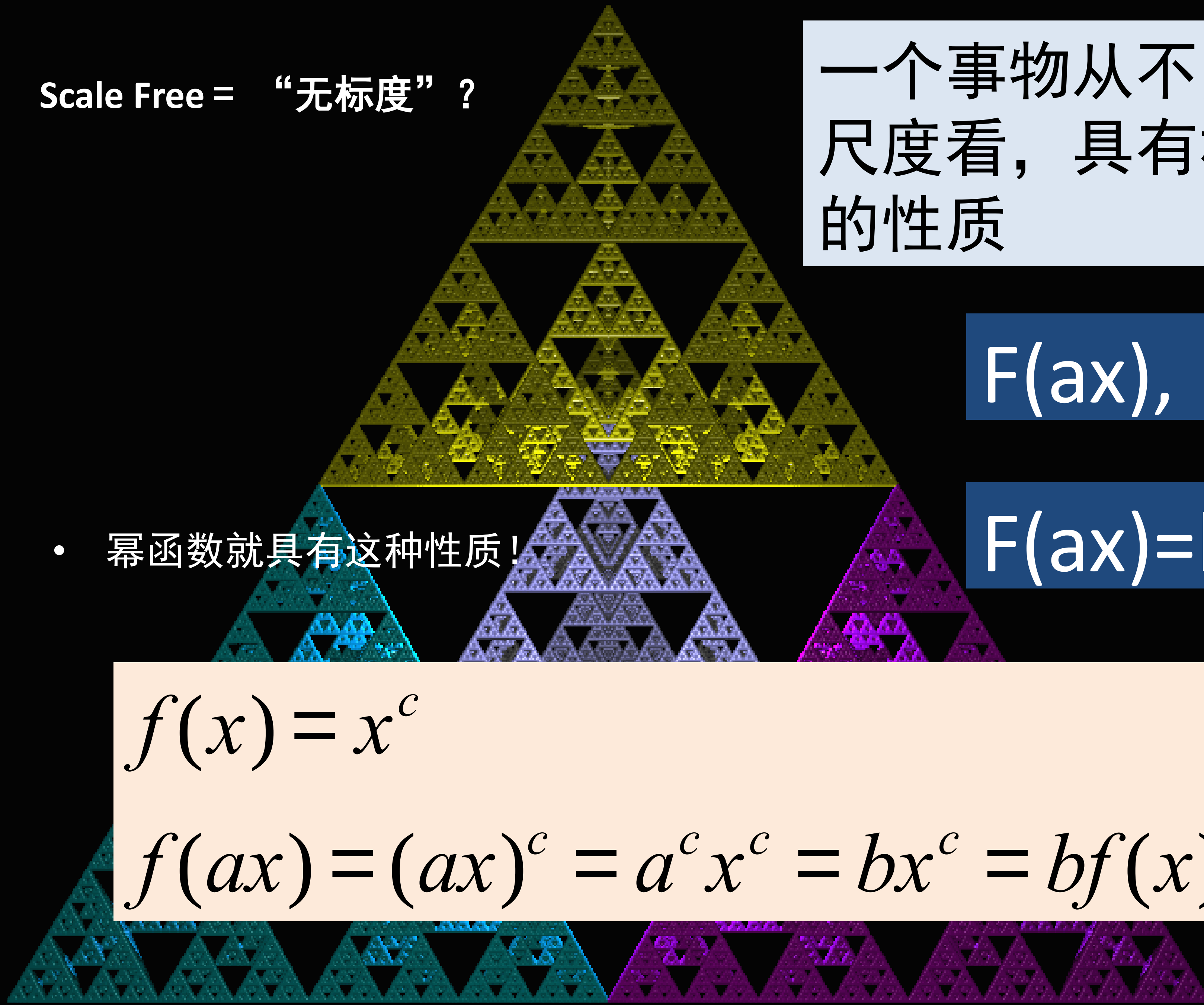
$$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$$



幂律的基本特性

- Scale free（不受尺度影响的）
  - Scale free函数隐含着自相似（self similarity）
- 平均行为不反映典型行为
  - “典型行为”—经常遇到的；
  - “平均行为”—总和／个数
  - 正态分布的“平均行为”反映“典型行为”
    - 典型看到“中等个子”，大个子很稀少

Scale Free = “无标度”？



一个事物从不同的尺度看，具有相同的性质

F(ax)，F(x)

F(ax)=bF(x)

- 幂函数就具有这种性质！

$$f(x)=x^{c}$$

$$f(ax)=(ax)^{c}=a^{c}x^{c}=bx^{c}=bf(x)$$

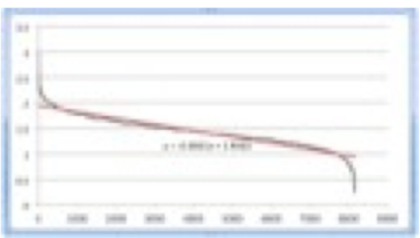
幂律的基本特性

- Scale free（不受尺度影响的）
  - Scale free函数隐含着自相似（self similarity）
- 平均行为不反映典型行为
  - “典型行为”一经常遇到的；
  - “平均行为”一总和 / 个数
  - 正态分布的“平均行为”反映“典型行为”
    - 典型看到“中等个子”，特别矮
      少

幂律分布比较容易看到“个大的”

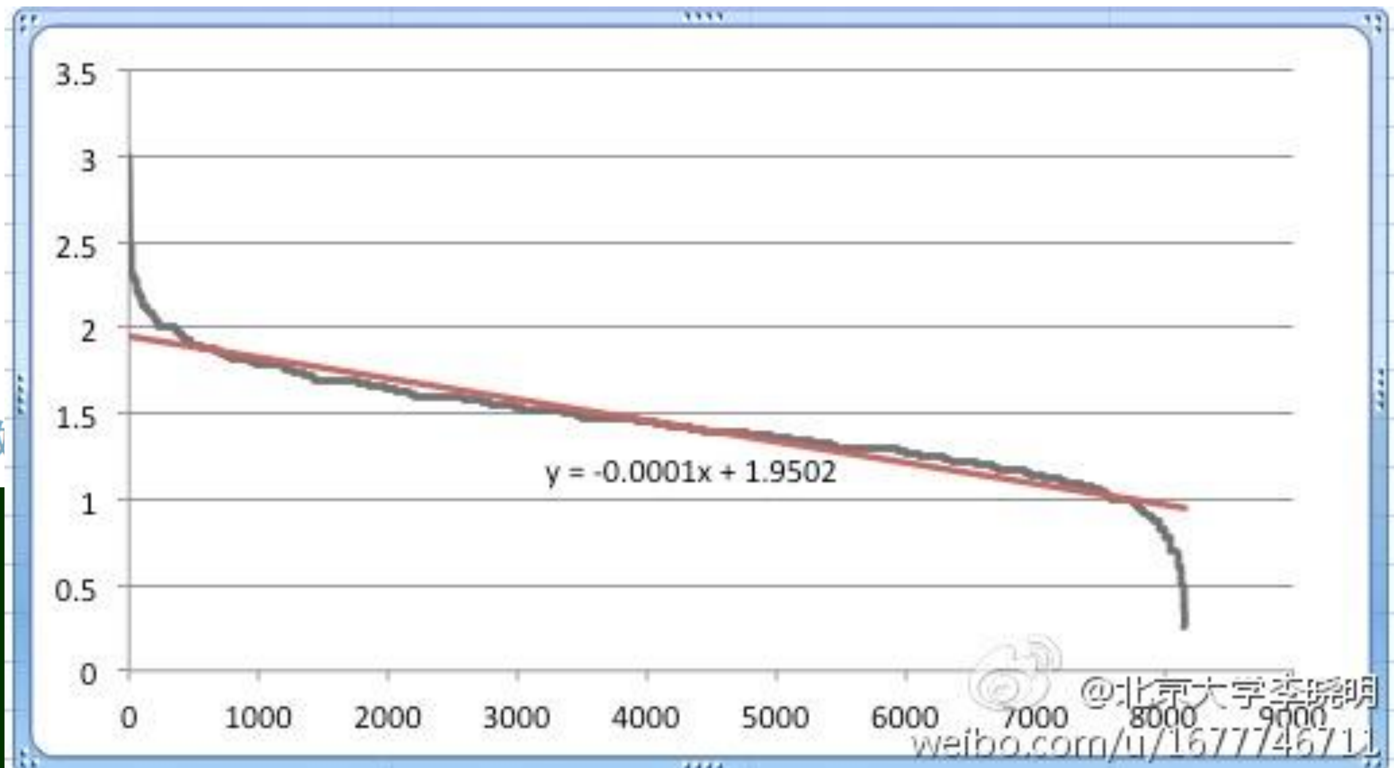
中国人均住房面积：符合幂律分布

前些时，北大的社会科学调查中心发布了《中国民生发展报告2012》，其中人均住房面积数引起了一些质疑。我昨天将那8000多个精心抽样的原始数据看了看，从高到低排序观察，十分典型的齐普夫现象（见下图），因此用平均的概念不妥，我看了一下中位数，27.5，比较靠谱。



+加标签

8月16日08:53　来自新浪微



$$f(x)=\frac{a}{x^2}=ax^{-2},\ x\in[1,n]$$

体会“典型”不同于“平均”的算例

To determine the normalizing factor a, set

$$\int_1^n f(x)\,dx=1,\ \text{i.e}$$

$$\int_1^n ax^{-2}\,dx=-ax^{-1}\Big|_1^n=a-an^{-1}=1$$

$$a=\frac{n}{n-1},$$ then, figure out the mean

$$\int_1^n xf(x)\,dx=\int_1^n ax^{-1}dx=a\ln x\Big|_1^n=a\ln n=\frac{n\ln n}{n-1}$$

suppose n=100, we have:

$$\frac{n\ln n}{n-1}=\frac{200\ln 10}{99}\approx\frac{200\times 2.3}{99}=4.65$$

see the probability observing larger than mean

$$-ax^{-1}\Big|_{4.65}^{100}=\frac{100}{99}\times\left(\frac{1}{4.65}-\frac{1}{100}\right)\approx 0.207,\ \text{also}$$

$$-ax^{-1}\Big|_{9.3}^{100}=\frac{100}{99}\times\left(\frac{1}{9.3}-\frac{1}{100}\right)\approx 0.1$$

取值范围

n=1, . . . , 100

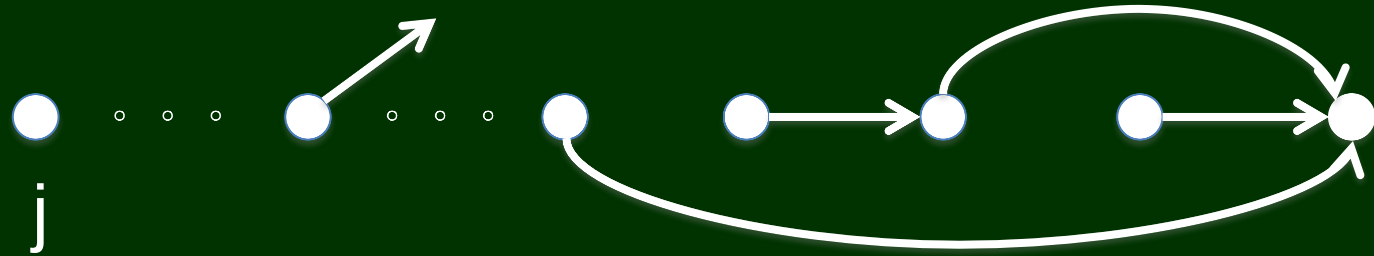
均值＝4. 65，相对比较小

意味着：看到比均值大的对象的可能性很高

具体算出来，看到较大对象的概率约为 0. 2

最后这个计算表明看到比均值大一倍对象的概率约为 0. 1

幂律的成因（“富者更富”模型）



- 网页按照顺序创建：1，2，3，…，j，…
- 当创建网页j 时，以概率p或1-p选择如下(a)或(b)执行

（a）以概率 p，均匀地、随机地选择一个早先创建的网页 i，建立一个从 j 到 i 的链接

（b）以1-p的概率，均匀地、随机地选择一个早先创建的网

此模型产生幂律ak^{-c}，其中的指数 c 取决于概率 p

为什么说这体现了“富者更富”

- 网页按照顺序创建：1，2，3，…，j，…
- 当创建网页j 时，以概率p或$1-p$选择如下(a)或(b)执行

（a）以概率 p，均匀地、随机地选择一个早先创建的网页 i，建立一个从 j 到 i 的链接

（b）以$1-p$的概率，均匀地、随机地选择一个早先创建的网页 i ，建立一个从 j 到 i 所指向的网页的链接

- 等价于说：
- ……

（b）以$1-p$的概率，按照已有入度成比例的概率，选择一个早先创建的网页 i，建立一个从 j 到 i 的链接。

富者更富效应的不可预测性

- “富者更富”也具有级联的意味，现实生活中有不少体现这种情形的现象
- 最初阶段充满不确定性，“富”到一定程度后就开始“起飞”
  - 与《哈利波特》同样质量的小说在同一时期其实很多，但真正流行起来的很少
  - 同样水平的歌星在同一时期其实很多，但真正出名的很少
- 一类事物流行史的细节不可能重演，但历史的结果宏观上总是如此（流行的分布）

历史平行演化的一次模拟实验

- 建一个音乐下载网站，向网民提供48首人们不太熟悉的歌曲的下载
- 该网站也公布每首歌曲的“已下载次数”，后面上来的人能够看到（从而就有一种促进富者更富的功效）
- 观察一段时间后那些歌曲下载量的分布

实验设计的妙处在：人们不知道他们被随机分到8个类似的网站之一（歌曲相同，初始状态相同）！

于是：研究人员看到了8段平行发展的历史。

与“长尾”（long tail）的关系

- 一类产品（例如书籍，个人音乐专辑）各个品种的销售量（流行度）常符合幂律

$$f(x) = \frac{a}{x^c}, \quad c \geq 2$$

发现销量为x的品种的概率

- 商业上人们更方便直接谈销量（而不是概率），设该类产品的品种总数为n，于是

$$n \times f(x) = \frac{n \times a}{x^c}, \quad c \geq 2$$

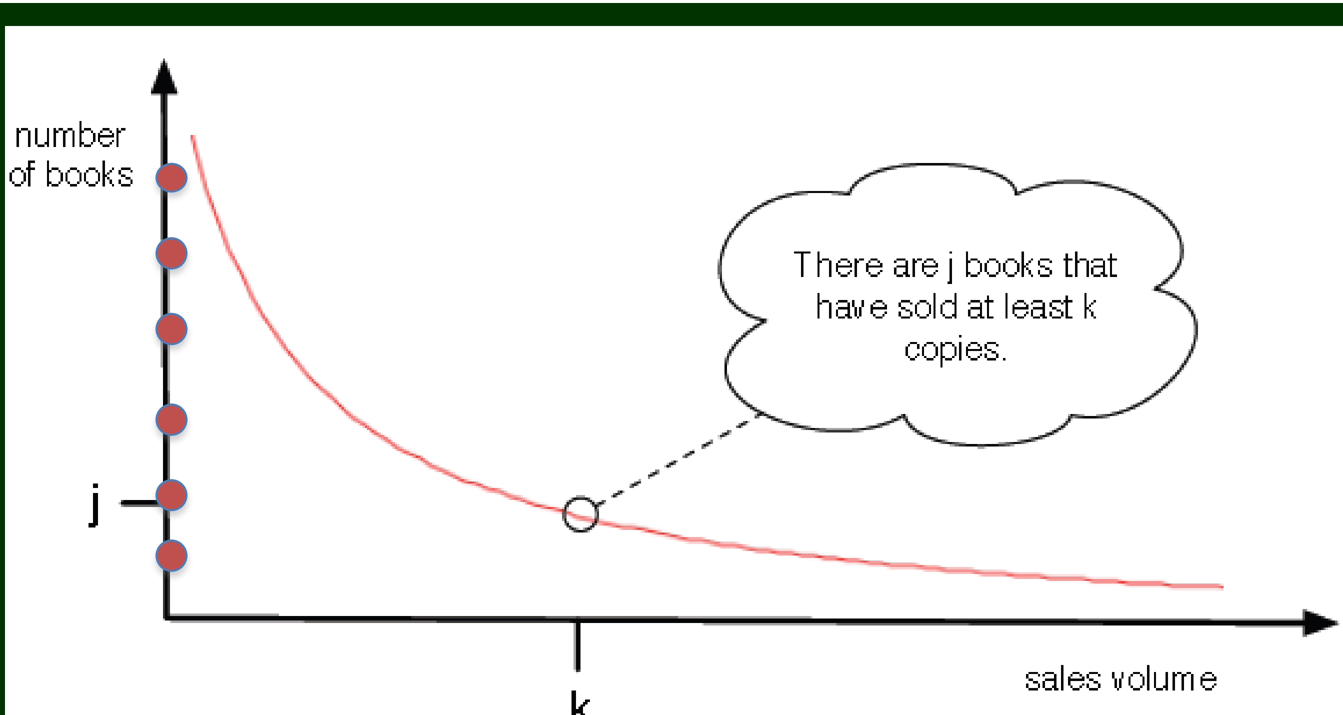
即销量为x的品种的个数

“长尾”（进一步）

也是幂函数
（但幂次变了）

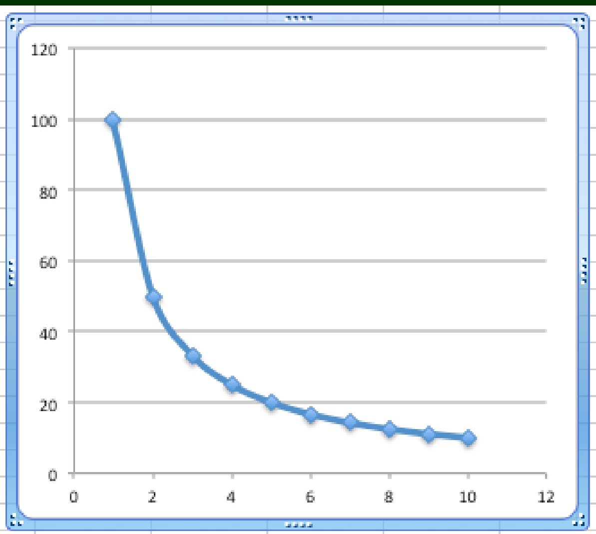
- 关心“销量至少为k的品种数”

$$\int_k^{\infty} n\cdot f(x)dx = \int_k^{\infty} \frac{n\cdot a}{x^c}dx = -\frac{n\cdot a\cdot x^{-c+1}}{c-1}\Bigg|_K^{\infty} = \frac{na/(c-1)}{k^{c-1}}, \quad c \geq 2$$

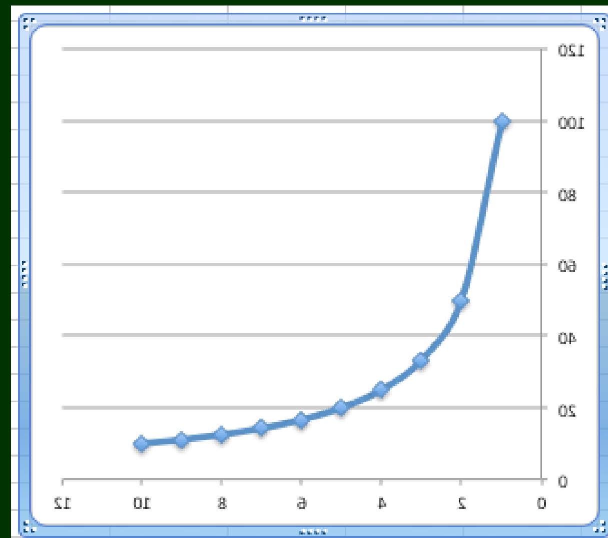


长尾的典型图示。由于降了一个幂次，尾巴显得更加明显

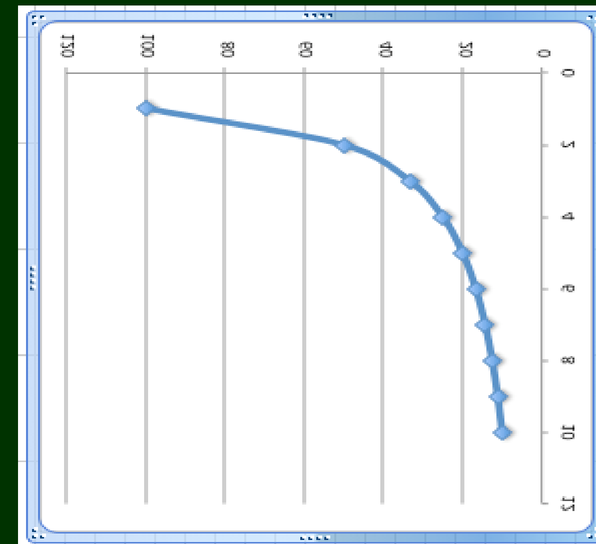
齐普夫定律（Zipf's Law）
——另一个视角看“长尾”



销量至少为k的品种数



“向左翻转”



“顺时针旋转”

- 横轴此时可看成“销量排名位次”，纵轴则是对应位次的销量。从函数关系看：

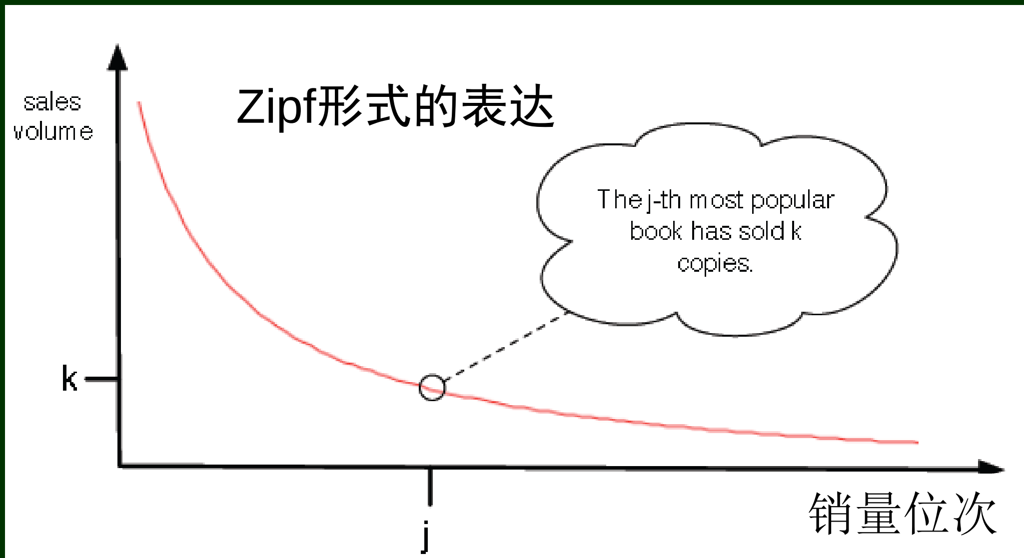
也是幂函数，尾巴更粗

$$y=\frac{a}{x^{c}},\quad c \geq 1$$

$$x^{c}=\frac{a}{y},\quad c \geq 1$$

$$x=\frac{a^{1/c}}{y^{1/c}}=\frac{b}{y^{d}},\quad d \leq 1$$

长尾效应与营销策略



结论是：如果一类商品的品种流行性分布为幂律，且品种足够多（即max很大），经营利基产品也能获得很大利益

$x=\dfrac{b}{y^{d}}$, consider all "non hits" sales 考虑top-100之后

assume d=$\dfrac{1}{2}$, corresponding to LT power=2

$$\int_{100}^{\max} b\cdot y^{-d}\,dy=-\frac{by^{-(d-1)}}{d-1}\Bigg|_{100}^{\max}=2b(\sqrt{\max}-10)$$

对应概率意义幂律中的幂次3

if d=1, corresponding to long tail power=1

$$\int_{100}^{\max} b\cdot y^{-d}\,dy=b\ln y\Big|_{100}^{\max}=b(\ln(\max)-\ln 100)$$

对应概率意义幂律中的幂次2

但有两个前提
＊降低库存成本
＊让顾客容易发现那些产品

“长尾”—“2 / 8律”

- “销量排前20%的书的销量之和占总销量的80%”，“少数人的财富之和占所有人财富之和的大部分”，…
- 设共有1000种书，销量满足齐普夫律，$y=b/x$
- 我们来看看排名前 20％的销量之和占总销量的百分比，也就是

$$\int_1^{200}\frac{b}{x}dx=b\ln x\Big|_1^{200}=b\ln(200)=5.3b$$

总销量为
b*ln(1000)= 6.9b
有5.3/6.9=0.77=77%

销售排行版、推荐、搜索

- 是促进“畅销产品”还是促进“利基产品”的销售？
- 排行版：推动富者更富
- 推荐（相关推荐）
  - 取决于“相关”的含义，若是“买了这产品的其他人通常也买了…”，则倾向于是富者更富；若是按照某种“内容相关性”，则可起到推动利基产品销售的作用
- 搜索：也是有两面性

富者更富过程的确定性近似

- 第一，在步骤t>=j,节点j的链入数是一个随机变量Xj(t)，其中Xj(t)有两个特点。
  - a)初始条件。因为节点j在最初的步骤j被创建时没有链入链接，因此Xj(j)=0。
  - b) Xj随时间的预期变化。当且仅当一个新创建的节点t+1直接链接到节点j时，节点j在步骤t+1后增加一个链入数。这种情况发生的概率是多少？节点t+1以概率p均匀随机地选择一个较早创建的节点并链接到该节点，以概率1-p选择一个较早创建的节点，并以这个节点的链入数成正比的概率创建到该节点的链接。前一种情况，节点t+1链接到节点j的概率为1/t。对于后一种情况，我们观察到在节点t+1被创建时，网络中链接的总数为t(每一个节点产生一个链接)，其中，有Xj(t)个链接指向节点j。因此，对于后一种情况，节点t+1链接到节点j的概率为Xj(t)/t。进而，节点t+1链接到节点j的总概率为：$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$
- 运行时间需要从0连续变化到N,用连续时间函数Xj(t)近似地替代节点j的链入链接数Xj(t)
  - 初始条件。因为Xj(j)=0，同样定义xj(j)=0
  - 微分方程确定这个增长速度：$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$

处理确定性近似

- 设q=1-p, 微分方程化为：$\frac{dx_j}{dt} = \frac{p+qx_j}{t}$ ，积分得到：$\ln(p+qx_j) = q\ln t + c$
- 对于一个常数指数c, 设 $A = e^c$, 则 $p+qx_j = At^q$ ，所以说 $x_j(t) = \frac{1}{q}(At^q - p)$
  根据xj(j)=0，得到 $A = p/j^q$ ，所以说 $x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}\cdot t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$
- 问题：对于一个给定值k，和 一个时间值t'那么在时刻t'有多少比例的节点拥有至少k个链人链接数？

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k \Rightarrow j \leq t\left[\frac{q}{p}\cdot k + 1\right]^{-1/q} \Rightarrow \frac{1}{t}\cdot t\left[\frac{q}{p}\cdot k + 1\right]^{-1/q} = \left[\frac{q}{p}\cdot k + 1\right]^{-1/q}$$

- 微分得到 $\frac{1}{q}\frac{q}{p}\left[\frac{q}{p}\cdot k + 1\right]^{-1-1/q}$ ，确定性模型预期链人链接数为k的节点比例与k-(1+1/q)成正比，这是一个幂律函数，指数为 $1+\frac{1}{q} = 1+\frac{1}{1-p}$.
- 对原模型的分析表明，以高概率随机生成的链接，链入链接数为k的节点比例确实与 $k^{-(1+1/(1-p))}$成正比。由确定性近似模型提供的启发式证明以一种简单的方式描述了这 个幂律指数1+1 /(1-p)是如何形成的。

要点小结

- 幂律（概率分布）是流行现象的主导规律
  - 但不是100%普适规律
  - “富者更富”是幂律的一种成因。发现一种流行现象的规律有意义，理解其成因更重要
- 符合幂律的流行现象也可以通过“长尾”或齐普夫定律来刻画
  - 它们本身也满足幂函数关系（但幂次不同）
  - 不仅幂律是“长尾”，还有其他长尾分布
- 对营销策略的启示