# 幂率与富者更富现象

## 及其与长尾、齐普夫定律等的关系

# 流行度（popularity）

- 同一类事物的不同实例被关注（认知、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？

# 以Web上网页得到的链接数量为例

- 得到一个链接，意味着得到某种“认可”。于是可认为得到的链接越多，流行度越高
- 问
  - 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？
  - 这个概率函数是否反映了一种规律，即普适于任何大规模搜集的网页集合？
  - 如果这是反映网页集合的一种规律，为什么会有这规律？它是否也适合其他具有流行性的事物？

# 先看如何回答第一个问题

- 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

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

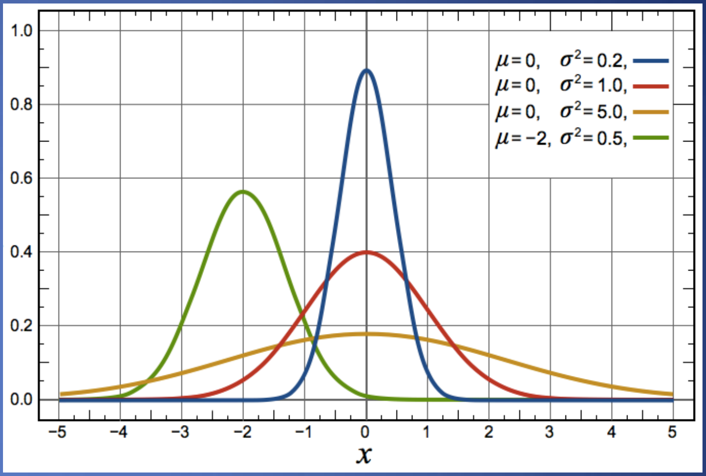
什么性质？
曲线是什么形状？

# 正态分布—随机量的一种规律

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

概率密度函数

μ: 均值；σ²: 方差；σ: 标准差



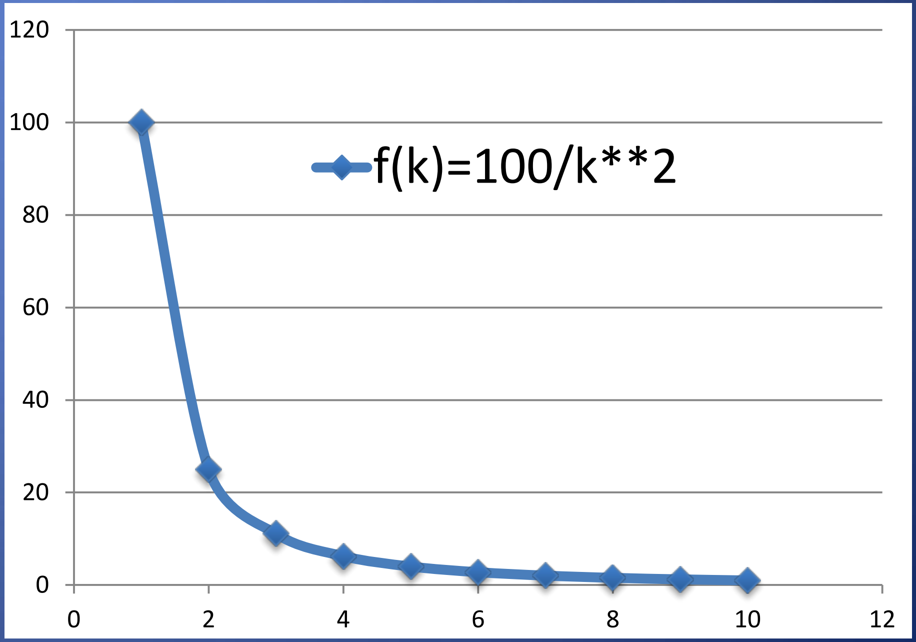

**中心极限定理**：大量独立同分布的随机变量之和（均值）是正态分布的随机变量；与原始分布是什么无关。

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：$f(k)=\dfrac{a}{k^c}=a\cdot k^{-c}$

- 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——“幂率”

| 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 |



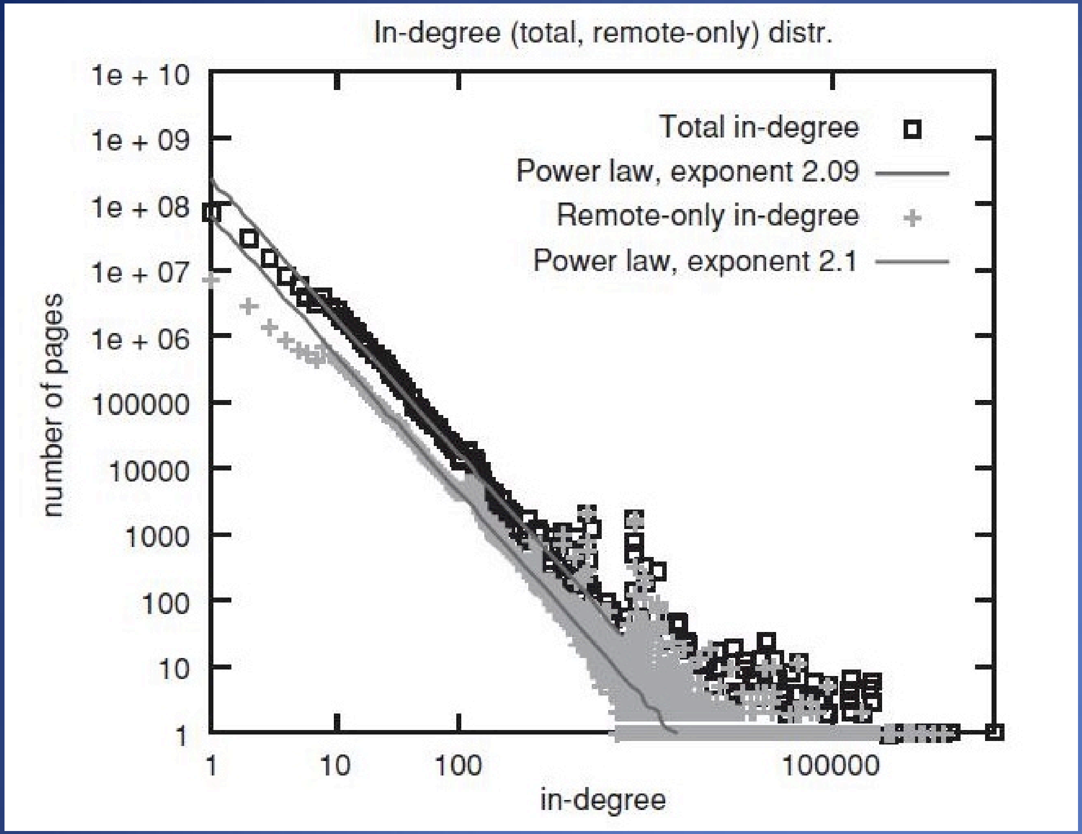
# 幂率的习惯（图形）表示

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

- 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
  - 以 k 为指数标度的横轴，f(k) 为指数标度的纵轴的图像是一条直线

1 2 3 4 … log(k)

$10^1$ $10^2$ $10^3$ $10^4$ … k




因此，给定一组原始数据

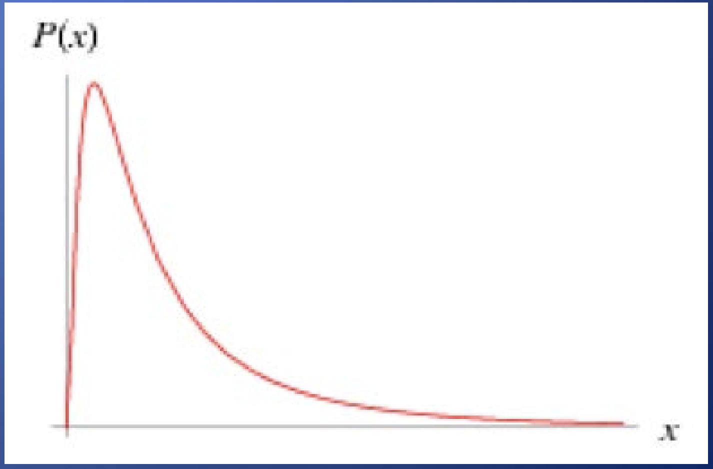
k: 1, 2, 3, …
f(k): …

- 为查看f(k)是否幂率，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。

# 幂率：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
- 一种书籍的销量
- …

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




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

# 幂率的基本特征

- Scale free（不受尺度影响的）
  - 数学上，一个函数f(x)称为“scale free”，若 **f(a*x)=b*f(x)**
  - 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

比较容易看到“个大的”

一个算例

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

To determine the normalizing factor a, set

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

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

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

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

suppose n=100, we have:

$$\int_1^n xf(x)dx=\frac{n\ln n}{n-1}=\frac{200\ln 10}{99}\approx\frac{200\times 2.3}{99}=4.65$$

see the probability observing larger than mean

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

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

取值范围

n=1, . . . , 100

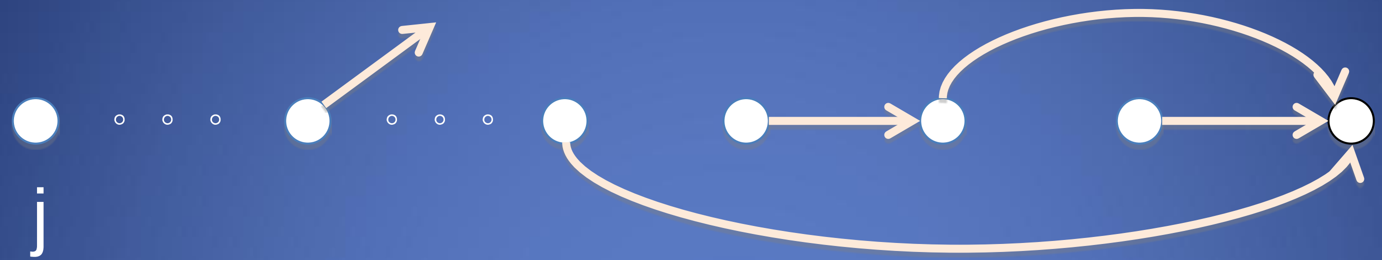
均值＝4. 65，相对比较小

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

# 幂率的成因（“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页 i，建立一个从 j 到 i 所指向的网页的链接。

此模型产生幂率，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：
- ……

  （b）以1-p的概率，按照与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。

# BA无标度网络模型

## Barabasi-Albert(B-A)模型

无尺度网络形成的两个基本机制：

(1)增长。

(2)优先连接。

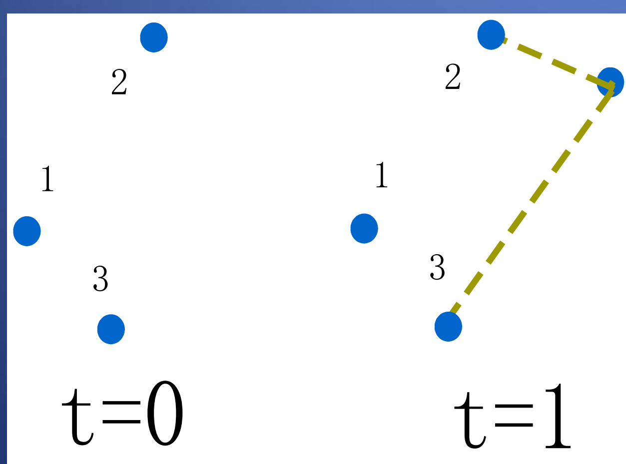
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}$$

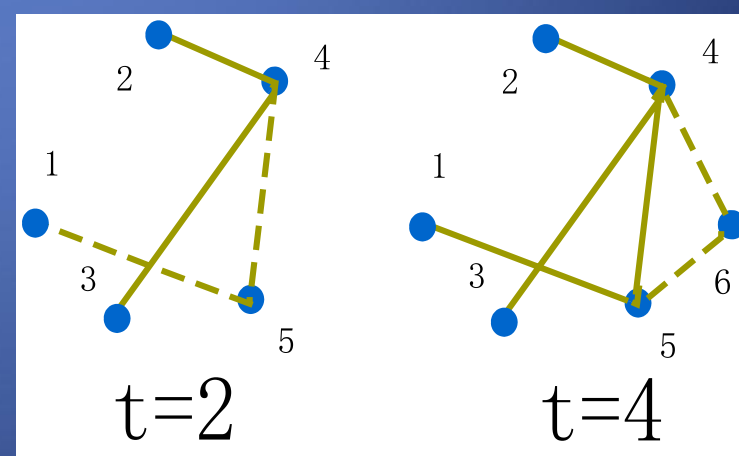
# B-A 模型的构建

$$\Pi(k_i)=\frac{k_i}{\sum_j k_j}$$

（1）增长：在初始时刻，假定系统中已有$m_0$个点，在以后的每一个时间步长中，我们新增一个度为m的点（$m<=m_0$），这m条边连向网络中已经存在的m个不同的点。

(2) 优先连接：当我们在原来网络中选择一些点被新增加的边连结时，这些点被连结的概率与这些点自身的度的大小成正比。比如度为$K_i$的点i 被新增点连结的概率为：



# 适应度模型

**BA模型中老节点具有较高的度，增加适应度**

(1) 增长：在初始时刻，假定系统中已有$m_0$个点，在以后的每一个时间步长中，我们新增一个度为m的点($m<=m_0$)，这m条边连向网络中已经存在的m个不同的点。

(2) 优先连接：

$$\Pi(k_i)=\frac{\eta_i k_i}{\sum_j \eta_j k_j}$$

# 局域世界演化模型

## 增加局域世界

(1) **增长**：在初始时刻，假定系统中已有$m_0$个点，在以后的每一个时间步长中，我们新增一个度为m的点($m<=m_0$)，

(2) **局域世界优先连接**：随机从网络已有的节点中选取M个节点(M>=m)，作为新加入节点的局域世界。第t步新加入的节点根据优先连接概率

$$\Pi_{\text{Local}}(k_i)=\frac{M}{m_0+t}\frac{k_i}{\sum_j \text{Local}\, k_j}$$

# 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

# “长尾”（long tail）又是什么？

- 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）常符合幂率

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

发现销量为x的品种的概率

- 人们更方便直接谈**销量**（而不是概率），设该类产品的总销量为n，于是

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

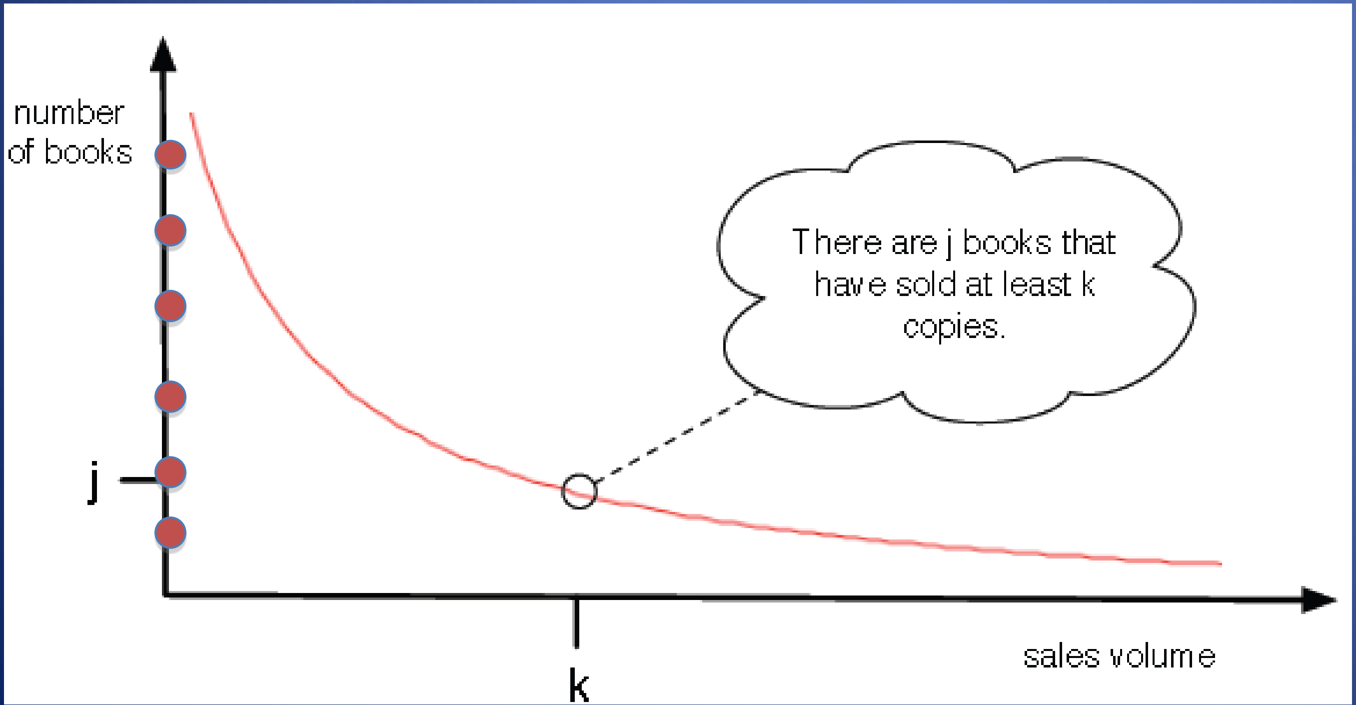
销量为k的品种的个数

# “长尾”（进一步）

也是“幂率”
（但幂次变了）

- 关心“销量至少为K的品种数”

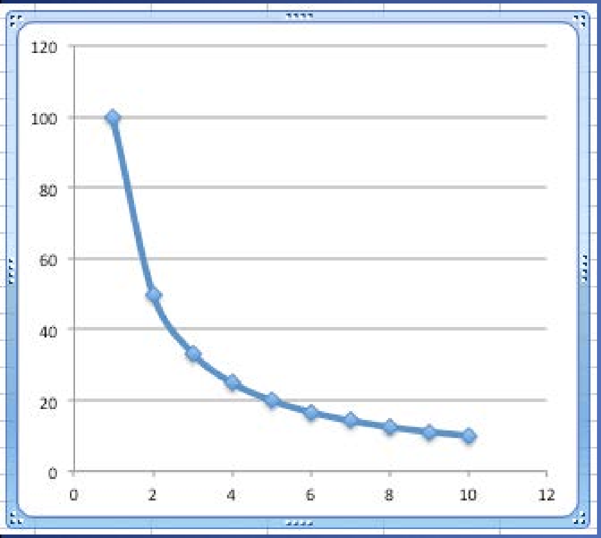
$$\int_K^\infty n\cdot f(k)dk=\int_K^\infty \frac{n\cdot a}{k^c}dk=-\frac{n\cdot a\cdot k^{-c+1}}{c-1}\bigg|_K^\infty=\frac{na/(c-1)}{K^{c-1}},\quad c\geq 2$$



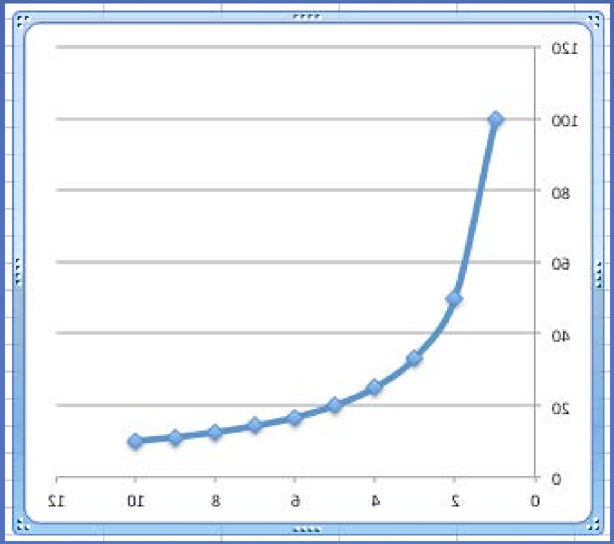

长尾的典型图示。由于降了一个幂次，尾巴显得更加明显

# 齐普夫定律（Zipf's Law）

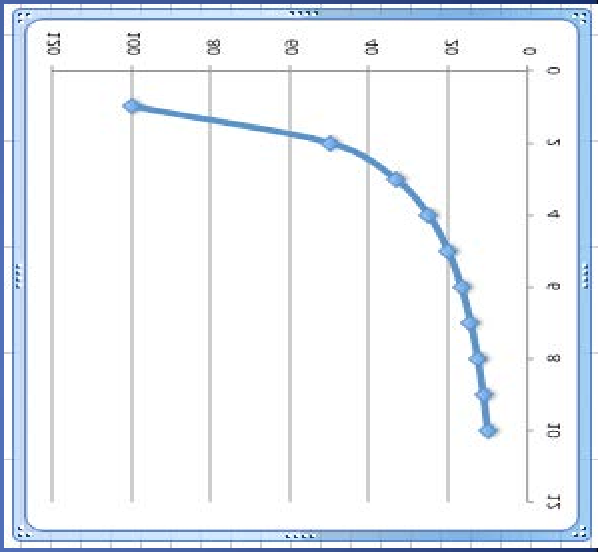
## ——另一个视角看“长尾”




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：

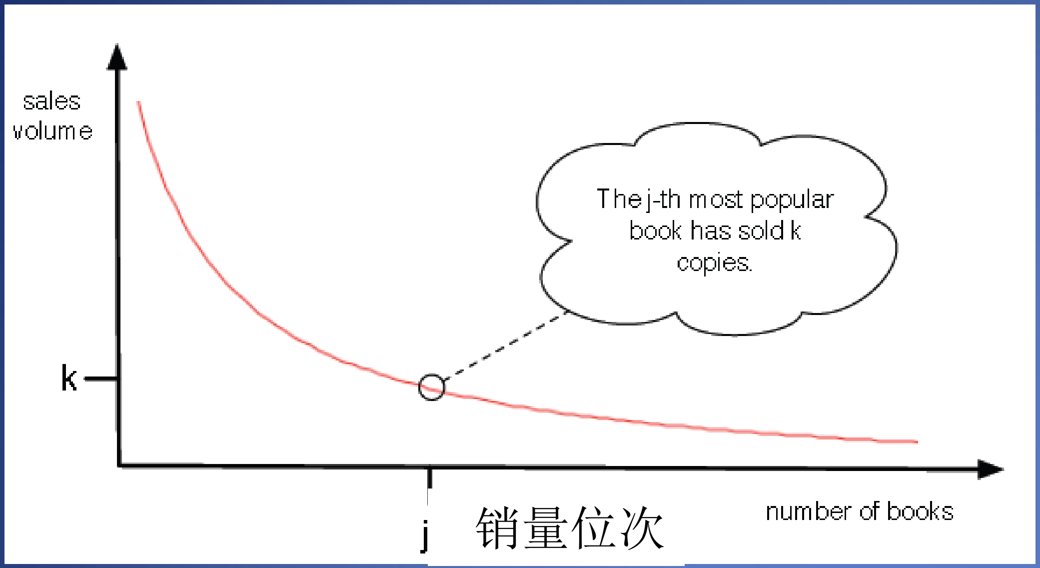
也是幂率！

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

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

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

# 长尾效应与营销策略




结论是：在长尾规律下，如果品种足够多（即max很大），经营利基产品也能够获得很大利益。

$x = \frac{b}{y^d}$, consider all "non hits" sales

考虑top-100之后

assume $d = \frac{1}{2}$, which correspond to c=2

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

对应概率意义幂率中的幂次3

if d=1, which correspond to c=1

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

对应概率意义幂率中的幂次2

但有两个前提

*降低库存成本

*让顾客容易发现那些产品

# 销售排行板、推荐、搜索

- 是促进“畅销产品”还是促进“利基产品”的销售？
- 排行板：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

# 要点小结

- 幂率是流行现象的主导规律
  - 但不是普适规律
  - “富者更富”是幂率的一种成因。发现一种流行现象的规律有意义，理解其成因更重要
- 符合幂率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）
  - 不仅幂率是“长尾”，还有许多长尾分布
- 对营销策略的启示

# 作业

- 第18章 2,3题