The Unified Neutral Theory of Biodiversity and Biogeography

读书会总结

第二次读书会
时间：2010 年 2 月 28 日下午 2 点
地点：三号会所小厅

第二章 物种相对多度相关理论回顾

胡中民

长期起来，物种多度分布是生态学的核心研究内容。同时，揭示植物丰度分布的内在机制对于保护生态系统多样性具有重要意义。本章就物种相对多度（Relative Species Abundance）的相关理论进行了回顾。

生态学研究通常有归纳（inductive）和演绎(deductive)两种途径。前者从所观察到的生态学现象出发，进行归纳总结，进而得到具有普遍意义的理论或模型；后者从一些基本假设出发，得出相关理论模型，再用观察到的现象对该理论模型进行验证。科学家对物种相对多度的研究也同样先后采用了这两种研究途径。

运用归纳法得出物种相对多度的代表性模型有：Fisher的对数级数分布模型（logarithmic series）和Preston的对数正态分布模型（lognormal distribution）。

对数级数分布模型

Fisher等(1943)的研究发现，在一个群落中，个体数多的物种很少，而个体数少的物种数却很多。基于获得的物种多度（即物种的个体数）数据，他们发现，物种的多度分布可以用对数级数（logarithmic series或logseries）分布模型表示（图1）。该模型的表达式为，若某物种的个体数为n（即x轴），那么与该物种具有相同个体数的物种数量为$\alpha x^n / n$（即y轴），其中α为表征群落多样性的指标,x为介于0与1之间的常数。用该模型计算可知，群落的物种数与个体数均是α的函数，而与其他因素无关，因而用α来度量群落的多样性得到了广泛的采用。

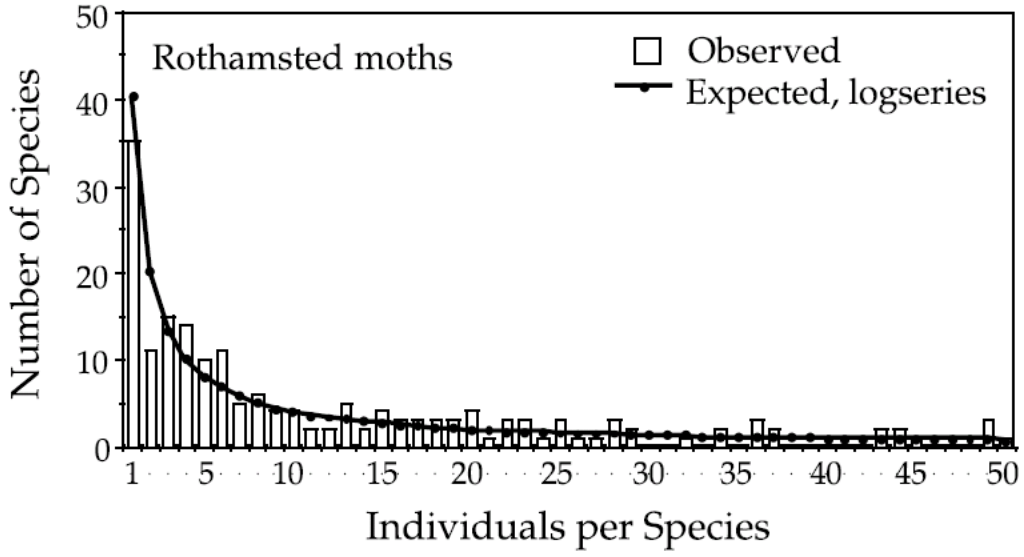


图1

对数正态分布模型

Preston (1948)的研究发现，Fisher的模型与他所观察到的现象并不吻合。Preston发现，个体数最少的物种并非最多，而是个体数位于中间的物种最多（图2）。后来许多研究发现该分布形式与观察到的现象更吻合。

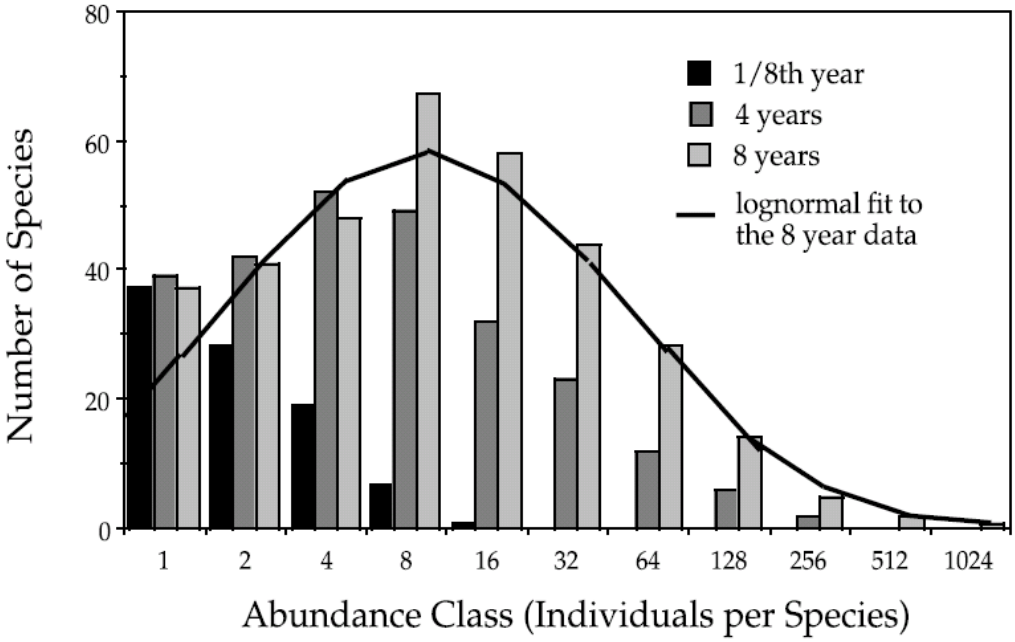


图2

在对图2的数据特征进行转换后，Preston得出了对数正态分布模型。他按

照物种的个体数分别为1、2～3、4～7、8～15 等进行分组（每组称为octave），再对个体数转换成以2为底的对数。进而得到图3形式的分布形式，即对数正态分布。

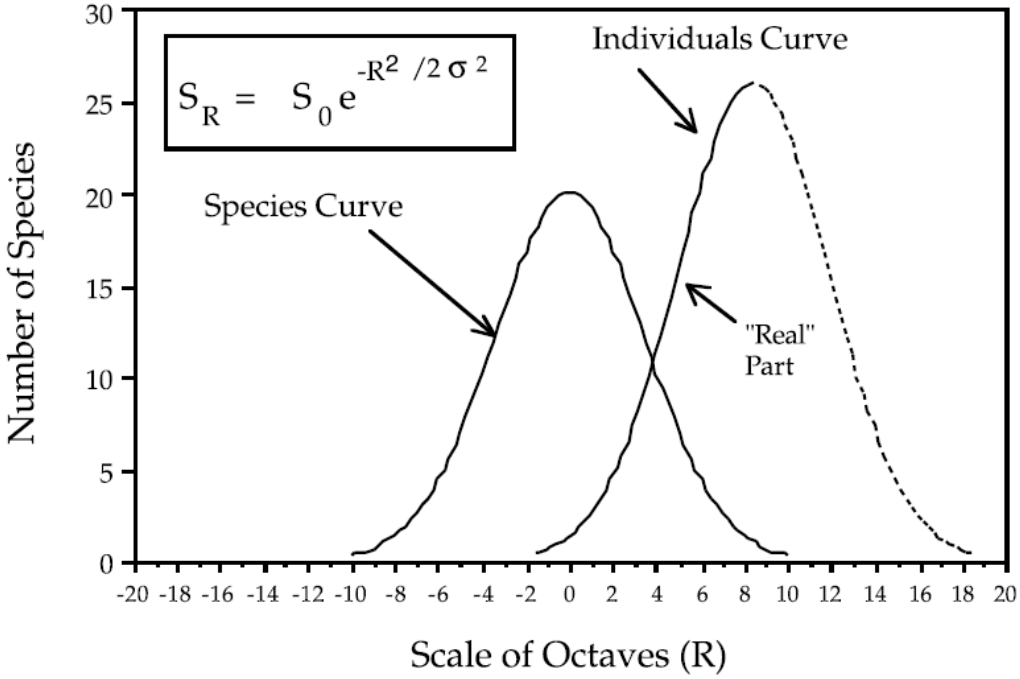


图3

Preston的另一重要发现是，物种多度分布与样本量的大小（sample size）密切相关，当样本量小时，分布格局可用Fisher的logseries分布表示，而当样本量大时，则用lognormal分布表示。进而他提出了一条“veil line”（图4），该样本量的大小使该veil line左右移动，从而出现不同的分布格局。

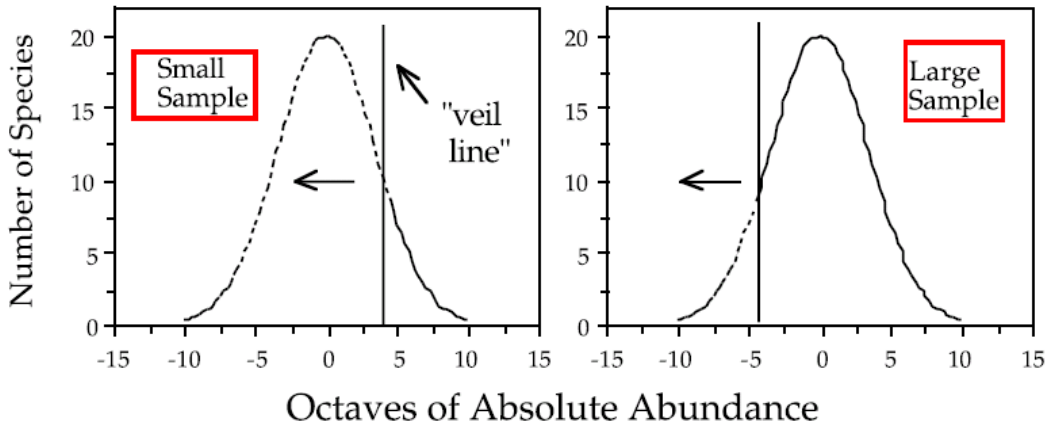


图4

Whittaker (1965)另一种方式表示了物种的多度分布，即优势度——多度分布曲线（dominance-diversity curve）（图5）。将物种的个体数由多到少排序，个体数最多的物种排x轴最左端，rank定为1，依次类推。Y轴对应物种的个体数

与群落总个体数的比，即相对多度。基于图5的表示方法，logseries表现为直线B，lognormal表现为曲线C。

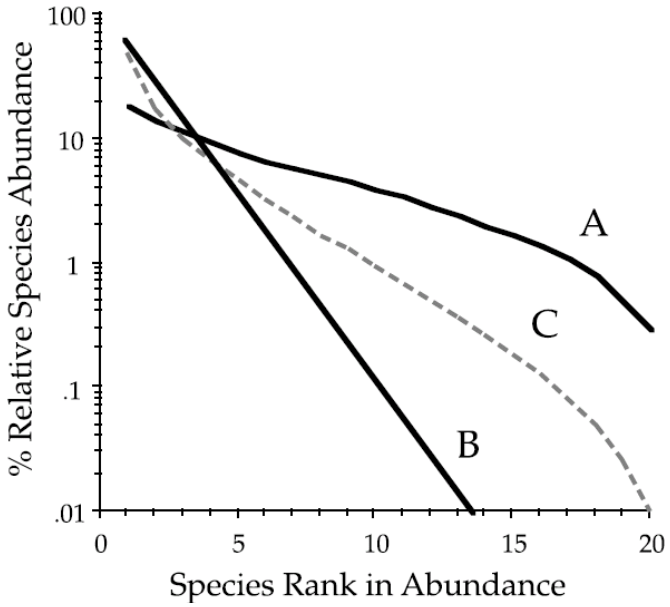


图5

随着研究的深入，人们的研究思路逐渐由归纳转向演绎。代表性理论有broken-stick理论，niche-preemption理论和sequential breakage理论。

broken-stick理论

MacArthur于1957年首先提出了“截棍子”理论(broken-stick hypothesis)。他假设群落中所有物种可利用的总资源有限，类似于一根长度一定的棍子，假设群落中的物种数为S，每个物种分别占有“棍子”的一小段，则随机将棍子截为S段，按截得的小棍由短到长排序（X轴），并与每段对应的物种数（Y轴）拟合，即得到图5中的曲线A。如果把图5的曲线C看作实际观测到的现象，显然A尽管表现出“S”型特征但显得太“平”。

niche-preemption理论

Whittaker (1965)提出了另一种资源分配假说：niche-preemption假说。与broken-stick理论不同的是，Whittaker假设群落中第一优势种占据群落总资源的比例为k(例如50%,也就是说棍子的50%被截去为第一优势种所占有),第二优势种再占据余下总资源的比例也为k,即剩下50%的50%，即25%的棍子再被截去为第二优势种占有，同样，第三优势种占12.5%,第四优势种占6.25%、。。。依次类推。同样，将截得的小棍（即被分配的资源）从长到短排序（X轴），即可得到图5的直线B。显然，该理论与logseries分布模型正好一致，但与曲线C却明显不符合。

sequential breakage理论

Sugihara (1980)在broken-stick理论的基础上提出了sequential breakage理论。Broken-stick理论是假设同时随机将棍子截为S段，而sequential

breakage理论则假设按顺序截棍子：（1）随机将棍子截为两段；（2）在这2段棍子中随机选其中的一段，再随机截成两段，此时共有3段棍子；（3）在这3段棍子随机选其中的一段，再随机截成两段，此时共有4段棍子；（4）在这4段棍子随机选其中的一段，再随机截成两段，此时共有5段棍子。。。依次进行。基于这种截棍子的方法，即可得到图5曲线C的分布形态，即与lognormal分布一致。由于碎石机粉碎石头的过程与该理论相同，Sugihara发现经碎石机粉碎后的小石块的大小——数量分布也符合lognormal分布。
尽管上述三个理论或多或少与观察到的数据相吻合，但都存在多严重缺陷。例如，对于broken-stick理论，是什么竞争机制导致资源随机被分割不明确，另外群落的物种数量也无法用该理论来预测；对于niche-preemption理论，假定固定值k难以用生物学机理来解释；对于sequential breakage理论，尽管其预测的分布与实际观察到的现象最吻合，但是其“截棍子”的方式还难以从生物学角度来解释，另外，“截”到何时是个头该理论未能指出，因而也无法预测群落的物种数量。再有，三种理论都未明确如何在现实群落中采样进而对其进行验证。最重要的是，这些理论都是静态的，没有考虑个体的出生，死亡和迁移过程，因而没有与经典的MacArthur and Wilson岛屿理论结合。

近年来，随着观测数据的不断丰富，人们发现，物种多度分布格局也并非与Preston的lognormal分布完全一致，而是在稀有种一端表现出更长的“尾”（图6），因而需要新的理论来囊括所有观测到的现象，预测物种数，与岛屿理论相结合。于是，中性理论来了。。。

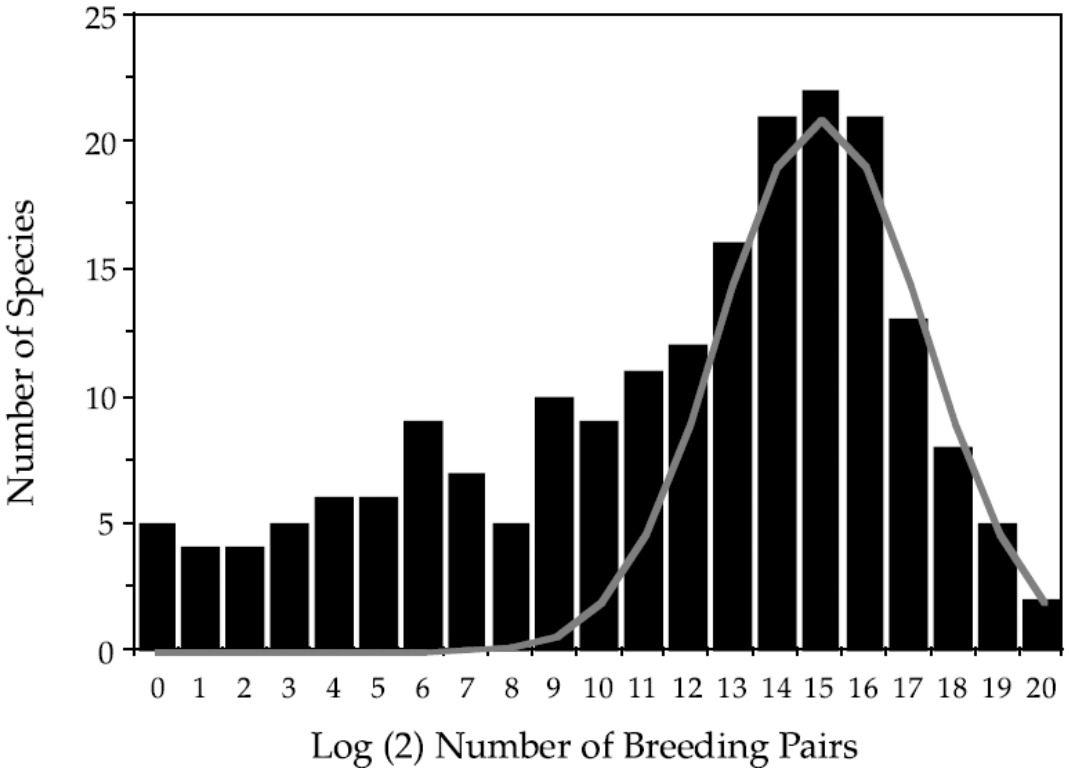


图6

The Unified Neutral Theory of Biodiversity and Biogeography

第二章 Jake 补充内容

一、概率分布、多度-Rank 曲线以及幂律

1、累积分布函数（CDF）和概率密度（PDF）

顺着王老师给我们上的“元宵节版的随机过程”课，继续澄清一些基本概念。首先，我们为什么要引入随机变量这个概念呢？那是因为我们首先可以从一种实数化的角度来看待随机事件了，比如说一个人的身高、年龄虽然都可以看作随机事件，但只有你把它们映射到实数空间，你才可以统一地看待这两个东西。另外，在实数随机变量的基础上，我们可以讨论一种叫做分布的概念。首先给定一个值域在实数集上面的随机变量 X，我们定义一个叫做分布函数（Cumulative distribution function, CDF）：

$$F(x) = P\{X \le x\}$$

注意我这个式子里面的 x 和 X 是有区别的。X 表示的是随机变量，而 x 表示的是一个数。因此这个式子的右端表示的是一个事件的概率，这个事件是：X 这个随机变量落到(-Infinity,x)这个区间上面的概率。这样，x 的变化不同，右边的概率大小也就不同，于是，我们就得到了一个关于 x 的函数，叫做 F(x)。

接下来，我们定义一种叫做概率密度的函数(Probability Density Function, PDF)：

$$f(x) = F'(x) = \frac{dF(x)}{dx}$$

当然，前提是 F(x)是可导的。

2、经验分布

这两个函数都是标准的概率论中的定义，大家似乎觉得太过稀松平常。但是，对这两个函数的理解很重要，一遇到实际问题，大家可能就会犯晕了。

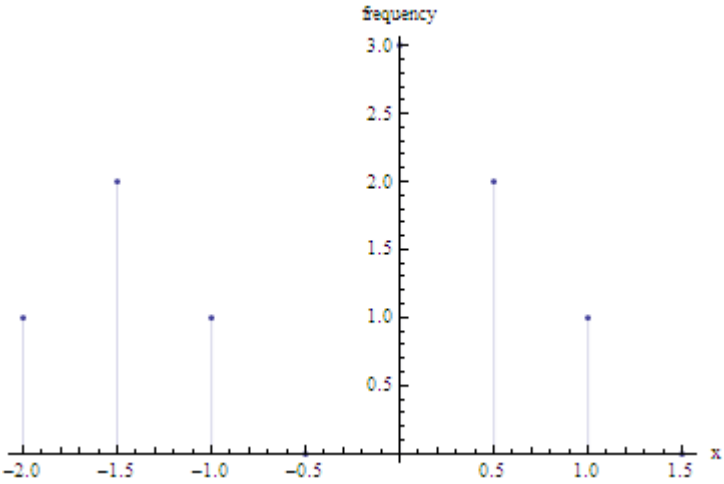
比如，考虑 10 个数构成的序列：

A= {1.30161, 0.0721849, -0.596532, 0.225591, 0.356764, -1.55749, -1.33701, -1.13029, 0.800888, 0.646423}

那么它的累积分布函数和概率密度函数是多少？

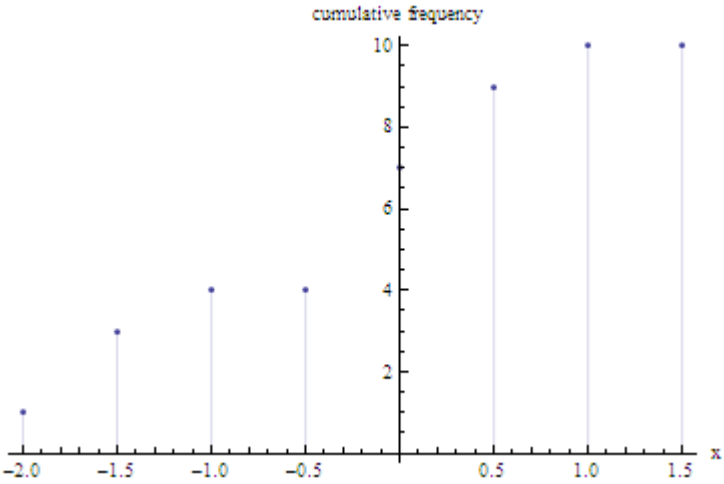
请注意，这个问题根本就没有答案！因为这种问法是错误的。所谓的 CDF 或者 PDF 只能针对随机变量来定义，而我们给出的是一个数列，它可能是某个随机变量 X 的 10 次的观测值（Outcome），但并不代表这个序列本身就是那个随机变量。

但是我们的确可以根据这个数列来猜测概率密度函数、累积概率密度函数。比如，根据这 10 个数，我们可以作这样一个柱状图：

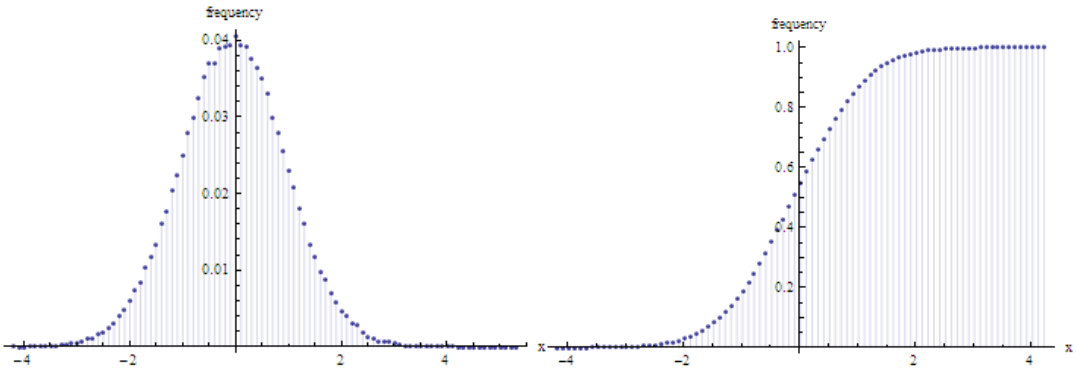


注意，这张图是这样做的：我们知道 A 序列中的 10 个数的范围都是在[-2,2]区间的，那么我们把这个区间分成 8 等份（每间隔 0.5），这样就得到了 8 个小区间。然后，我们计算 A 中 10 个数落到每一个小区间中的个数，我们就得到了：`{1,2,1,0,3,2,1,0}`。它表示落到[-2,-1.5]区间中的数有 1 个，落到[-1.5,-1]区间中的数有 2 个等等。这样把这些点画在图上就得到了上面的图。注意，图上每一个点的纵坐标表示 10 个数中落到该区间的个数，或者也可以理解为这 10 个数会访问该区间的频率次数，所以纵坐标也往往用频率表示。这张图的画法与 PDF 相似，只不过它来自实际数据而不是随机变量，所以我们称它为**经验概率密度分布图**。

还是这 10 个数，我们还有另外一种柱状图，还是把[-2,2]分成 8 个小区间，横坐标不变，但是纵坐标改为这 10 个数中有多少个是比区间右端点小的数？这样就得到了一种**经验累积密度**图：



我们看到，只要你给我拿出一列数，无论这些数服从什么概率分布，也不论这堆数究竟是不是一个随机变量的不同观测值，原则上讲我都能画出上面的两种图。但是，如果这个数列真的是某种随机变量的观测值的话，那么，随着数的增多，这两个图就会分别趋近于那个真实的随机数列的 PDF 和 CDF。例如，下列两图就是用 10000 个标准正态分布得到的经验分布图。



可以说，真正的随机变量分布我们在现实生活中根本就看不到，我们只能通过大量数据的积累而反过来推测随机变量的分布情况。

2、Rank 曲线

我们已经说过了，其实随机变量的分布函数根本就是不可测量的，因为我们只能得到随机变量所产生的观测值，然后通过大量的观测去反过来归纳随机变量的分布情况。除了用上面的经验分布图方法表示已知序列背后的统计规律外，我们还经常用一种我称之为 Rank 曲线的方法来表示一组给定的序列。因为这个 Rank 曲线很容易获得，而且本质上讲，当数据量非常大的时候，Rank 曲线就等价于累积分布曲线（CDF），所以像 Hubbell 书中的物种多度曲线、Zipf 律等等就都是用这种 Rank 曲线来说事情。

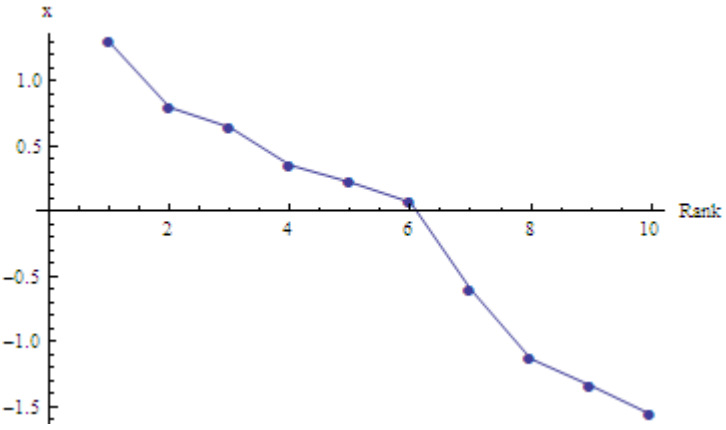
我们还用这个 A 序列来说明：

A= {1.30161, 0.0721849, -0.596532, 0.225591, 0.356764, -1.55749, -1.33701, -1.13029, 0.800888, 0.646423}

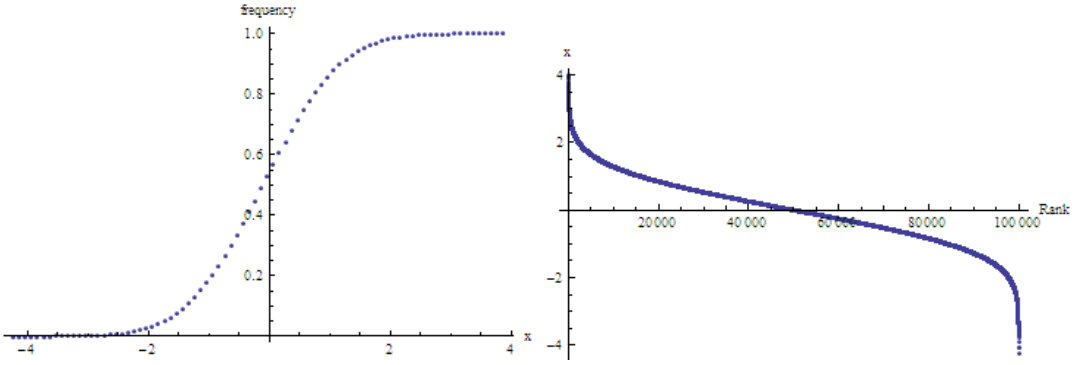
这次，我们不做任何处理，而是简单的将它们从大到小排序：

{1.30161, 0.800888, 0.646423, 0.356764, 0.225591, 0.0721849, -0.596532, -1.13029, -1.33701, -1.55749}

然后以 Rank（序号）为横坐标，以排在第 i 个位置的数为纵坐标就得到了这个 Rank 曲线：



当数据量大的时候，我们会发现其实这种 Rank 曲线会逼近于随机变量的累积概率函数(CDF)曲线的轴对称+平移，例如，我用计算机生成了 10000 个正态随机数，然后求它的经验累积分布曲线和 Rank 曲线：



其实，这两个曲线很相似，我们只要把 rank 曲线沿着 y=-x 这条对角线反射一下，再往上平移 1 个单位就几乎能得到左边的累积柱状图。当数据量非常庞大的时候，这种关系就会更加明显。这并不是一种巧合，因为本质上讲，这两条曲线就是一回事。我们可以用数学证明这个事实。

假设一个随机变量 X 的 N 次观测值的 Rank 曲线为函数：f(r)，其中 r 为排序值，f(r)表示排在第 r 个位置上的观测值。那么，我们不妨问，在这 N 个数中，大于 f(r)的数有多少个呢？我们知道，根据 Rank 曲线的定义，排在 r 左边的数都是比 f(r)大的，排在 r 右边的数都是比 f(r)小的，所以比 f(r)大的数就一共有 r 个。所占总数的比例是 r/N。

又根据大数定理，当观测值 N 趋近于无穷大的时候，频率就会趋近于概率，因此我们这里计算的比 f(r)大的数所占的比例 r/N 就相当于是概率。于是，我们就有这个式子：

$$P\{X > f(r)\} = r/N \qquad *$$

进一步，我们做变量替换，将等式左边的大括号里面的 f(r)去掉，即令：

$$y = f(r)$$

因为 f(r)一般来讲是一条单调的曲线，所以反函数存在，于是：

$$r = f^{-1}(y)$$

代入*式，就得到了：

$$P\{X > y\} = f^{-1}(y)/N$$

而根据概率运算法则，我们知道：

$$CDF(X) = F(y) = P\{X \le y\} = 1 - P\{X > y\} = 1 - f^{-1}(y)/N$$

所以，随机变量的累积概率分布函数就是：

$$F(y) = 1 - f^{-1}(y)/N \qquad **$$

进一步对这个式子两边对 y 求导，并按照反函数的求导法则，有 X 的概率密度函数(PDF)为：

$$PDF(X) = -\frac{1}{Nf'(f^{-1}(y))}$$

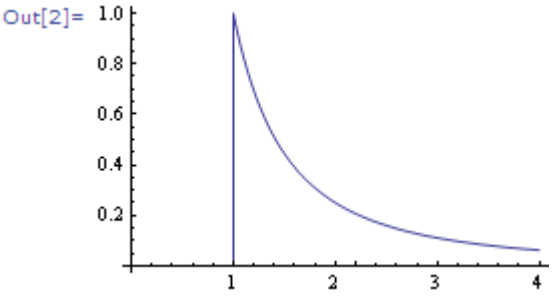
我们看到，只要 f 的具体形式确定，那么 X 随机变量的 CDF 和 PDF 也就全部确定了，所以这二者实际上是等价的。因此，考察物种的多度分布，我们既可以用概率密度函数，又可以用 rank 曲线，两者本质上是等价的。

3、幂律与 Zipf 律

作为练习，我们来考察一种特殊的概率分布，即幂律分布，在有些场合下也叫做 Pareto 分布，它具有下列的 PDF：

$$f(x)=\begin{cases}\alpha k^{\alpha}x^{-1-\alpha} & if\ x\geq k\\ 0 & if\ x<k\end{cases}$$

让 k=1，alpha=1，画出图来就是：



其中，k 表示幂律分布的起始范围，alpha 表示幂律指数，alpha 越大，那么这条曲线就会越陡。而当 alpha 小的时候，这条曲线就会拖一条长长的尾巴，因此具备我们所说的宽尾特征（fat tail）。

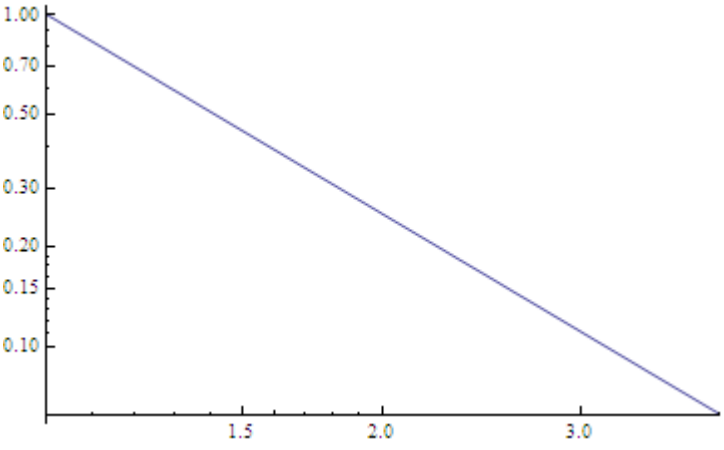
我们看到函数在 x>k 的部分是一个幂函数，也就是具备这样一种形式：

$$f(x)=ax^{-\alpha}$$

我们对上式两边取对数，得到：

$$\log(f(x))=\log(a)-\alpha\log(x)$$

我们看到，这个时候 log(f(x))这个变量和 log(x)这个变量就完全呈现出一种线性的关系，这会为我们的参数拟合等工作带来极大的方便，因此考察实际序列的经验分布的时候，我们会经常把数据点画在双对数坐标下，如果这些数据点构成一条直线，那么就说明这个分布是幂律的。比如，对上面的图画在双对数坐标下，得到：



其中，斜率就是-alpha。

进一步，我们考察幂律分布的 CDF，只要对 PDF 积分就可以了：

$$F(x)=\int_{k}^{x}\alpha k^{\alpha}t^{-1-\alpha}dt=1-(\frac{k}{x})^{\alpha}$$

根据上一节得到的**式，我们可以得到 Rank 曲线为：

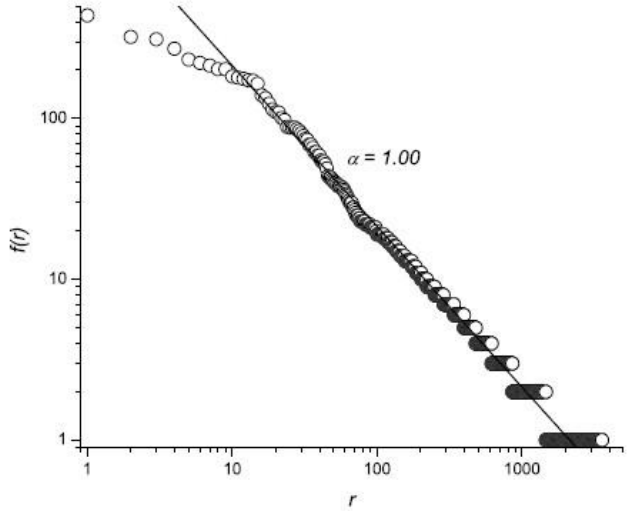
$$g^{-1}(x) = N(1-F(x)) = \frac{Nk^{\alpha}}{x^{\alpha}}$$

求反函数，得到：

$$g(r) = \frac{1}{N^{1/\alpha}k} \cdot r^{-1/\alpha}$$

我们发现，对于幂律分布，无论是累积分布曲线还是 Rank 图，都是幂律的形式，只不过它们的指数互为倒数。当我们把真实数据的 Rank 图也画在双对数坐标系下，也能得到一条直线。

对于很多实际数据，比如一篇文章中的不同单词的使用频率就可以按照从大到小的顺序排列，从而得到一张 Rank 图，这张图画在双对数下近似一条直线，这就被称之为 Zipf 律，因为是语言学家 Zipf 最早在英文中发现这个现象的，例如：



而且在通常情况下，直线的斜率刚好是-1。幂律有很多性质，我们可以考察一个幂律分布随机变量的均值：

$$m = \int_{k}^{\infty} \alpha k^{\alpha} x^{-1-\alpha} dx = \begin{cases} \dfrac{k\alpha}{\alpha - 1} & \alpha > 1 \\ +\infty, & \alpha <= 1 \end{cases}$$

也就是说 alpha<1 的时候，均值是无穷大。在现实社会中，很多变量的分布都是指数小于 1 的幂律分布，所以它们的均值是不存在的，这意味着，我们不能为这类随机变量找到一个特征值来作为整体的代表。例如，社会上的财富分布在高端就遵循幂律分布，这就意味着，你遇到多么富有的人的可能性都有，并且平均的财富根本就不存在的。（但在现实中，我们总能计算出一个确定的经验平均值）。

关于幂律和 Zipf 律的更多性质，以及解释性模型，请大家不要偏信旁言，最好是参考 Mark Newman 的一篇权威性文章 Power laws, Pareto distributions and Zipf's law:，如果你进一步想做一些关于幂律分布的实证数据工作，请参考 Aaron Clauset 的一篇经典文章 Power-law distributions in empirical data。另外，计算士为大家写了一个比较完善的关于幂律生成模型的总结。

二、切线段模型

第二章中有三个切线段模型比拟真实生态群落中物种相对多度的分布。下面，我们分别讨论：

1、模型 1

将一条单位长度的线段快刀斩乱麻，随机切成 N 多份，那么如果我们把这些小线段的长度作为一个随机变量 X，那么 X 的分布是什么呢？

估计某些人不动脑子的就会想，那显然是一个均匀分布吗！因为你完全均匀的去切这 N 刀。但是，我们这里问的不是这 N 刀切出来的坐标的分布，而是问切出来的线段长度的分布。所以，很明显这些线段长度不是均匀分布（从小到大等概率出现）。

第二种可能猜到的答案是正态分布！既然我切点是均匀的，那么得到的线段长度自然也是等间距的，所以肯定是大部分线段共享一个长度 1/N，而有少数的线段会对这个结果有所偏离，所以构成一个正态分布。

可惜的是，这个答案也是错的，首先一点，我们会看到如果是正态分布，并且中心在 1/N 处的话，那么随着 N 的增大，这个中心会趋近于 0，所以比 1/N 短的线段出现的概率就会趋近于 0，所以这就会遇到了一个边界效应问题，线段长度只能比 1/N 长，而不会比 1/N 短，所以结果会偏离正态分布。正确的答案应该是指数分布。要得到这个分布，我们分别用三种方法来求一求。

方法 1：最大熵方法

我们可以把线段看作分子，它们拥有的长度看作分子的能量，这样无论如何切，线段的总长度不变，也就意味着分子的总能量要守恒，所以求线段长度的分布就等价于求分子们经过碰撞之后导致的能量的分布。按照统计物理的玻尔兹曼模型，最后的分布可以用最大化熵的方法求出。

设线段长度刚好等于 x 的线段个数是 n(x)个，那么，我们显然有约束：

$$\int_0^1 n(x)dx = N$$

这就是说总共应该有 N 个小线段，还有：

$$\int_0^1 xn(x)dx = 1$$

这就表示所有线段的长度加起来应该为 1。在满足这两个约束条件下，我们求熵的最大化，即最大化下面的函数：

$$S(n(x)) = -\int_0^1 \frac{n(x)}{N}\log\frac{n(x)}{N}dx$$

注意，S 实际上一个关于函数 n(x)的泛函，所以我们可以对 n(x)进行变分从而得到 n(x)的形式（具体方法可以参见这个帖子：读文笔记）：

$$n(x) = \frac{e^{-Nx}}{e^N - 1}N^2 e^N,$$

因此，n(x)是一个负指数分布。

方法 2：曹旭东的贝叶斯概率法

我把曹旭东当年的解法放到这里，大家看按照他这种比较“暴力”的方法，也能推出负指数

分布。它这种方法虽然比较艰难，但是却能得到当不均匀撒点的时候，也能得到一般的公式。大家如果自信数学好的话，可以揣摩揣摩他的方法。原文见：这里。

【问题】

长度为 n 的直线，随机分布 n 个点，$X_1, X_2, \ldots, X_n \sim f(X)$，随机变量独立同分布，其中 $f(X) = \frac{1}{n}$ 表示均匀分布的概率密度函数。

证明在这 n 个点分割出的 n+1 条线段中随机抽样，选出的线段长度 L 是一个随机变量，证明当$n \to \infty$时随机变量 L 的概率密度满足均值为 1 的指数分布。

【证明】

首先给出一些定义：

$P(L \le t)$：线段长度 L 小于等于 t 的概率

$L_{i,j}$表示X_iX_j之间线段的长度。

下面给出一个非常重要的等式：

$$P(L \le t) = P(L_{i,j} \le t | X_i, X_j \text{相邻})$$

这个等式多少有些费解，我们继续往下走。

$$P(L \le t) = P(L_{i,j} \le t | X_i, X_j \text{相邻}) = \frac{P(L_{i,j} \le t,\ X_i, X_j \text{相邻})}{P(X_i, X_j \text{相邻})}$$

下面分别给出$P(L_{i,j} \le t,\ X_i, X_j \text{相邻})$和$P(X_i, X_j \text{相邻})$的具体表示。先从简单的开始：

$$P(X_i, X_j \text{相邻}) = \frac{2(n-1)!}{n!} = \frac{2}{n}$$

求$P(L_{i,j} \le \mathrm{t},\ \mathrm{X}_i, X_j$相邻$)$稍微复杂一点要用到全概率公式：

$$P(L_{i,j} \le \mathrm{t},\ \mathrm{X}_i, X_j\text{相邻}) = \int_{|X_i - X_j| \le t} \left(1 - \frac{|X_i - X_j|}{n}\right)^{n-2} \frac{1}{n^2} \mathrm{d}x$$

$$\text{其中} \mathrm{d}x = \mathrm{d}x_1\, \mathrm{d}x_2$$

令$\mu = -|X_i - X_j|$，则有$\mathrm{d}x = 2(n-\mu)\,\mathrm{d}\mu$，上面的等式可以化成如下形式：

$$P(L_{i,j} \le \mathrm{t},\ \mathrm{X}_i, X_j\text{相邻}) = \int_0^t \left(1 - \frac{\mu}{n}\right)^{n-2} \frac{2(n-\mu)}{n^2} \mathrm{d}\mu$$

综合以上各式有：

综合以上各式有：

$$P(L \le t) = P(L_{i,j} \le t \,|\, X_i, X_j\text{相邻}) = \frac{P(L_{i,j} \le \mathrm{t},\ \mathrm{X}_i, X_j\text{相邻})}{P(X_i, X_j\text{相邻})}$$

$$= \frac{\int_0^t \left(1 - \frac{\mu}{n}\right)^{n-2} \frac{2(n-\mu)}{n^2} \mathrm{d}\mu}{\frac{2}{n}} = \int_0^t \left(1 - \frac{\mu}{n}\right)^{n-2} \frac{(n-\mu)}{n} \mathrm{d}\mu$$

当$n \to \infty$时

$$P(L \le t) = \int_0^t e^{-\mu}\, \mathrm{d}\mu$$

概率密度函数为：

$$g(t) = \frac{\partial P(L \le t)}{\partial t} = e^{-t}$$

也就是均值为 1 指数分布。

如果$X_1, X_2, \ldots, X_n \sim f(X)$，分布函数$f(X)$不是均匀分布，整个求解过程中需要变化的是：

$$P(L_{i,j} \le t,\ X_i, X_j \text{相邻}) = \int_{|X_i - X_j| \le t} \left(1 - \int_{X_i}^{X_j} f(x)\,dx\right)^{n-2} f(X_i) f(X_j)\,dx$$

虽然能够写出公式，但是对于非均匀分布的情形，这个式子是困难的。

方法 3：随机点过程法

这种方法就是王钺所说的随机过程法。具体是这样的，假设我们已经随机把切点撒到了线段上，但是线段还没有断开。另外，假设你是一只小蚂蚁沿着线段的左端往右端爬，如果你爬的步伐很小，使得每一步要么遇到切点，要么不会遇到切点。我们设切点撒到线段上的密度为ρ 。在一个时间步中，你遇到切点的概率是ρ ，遇不到切点的概率是 1-ρ ，这样，在经过了 T 个时间步之后，你遇到了 k 个切点的概率就是：

$$P(k) = C_T^k \rho^k (1-\rho)^{T-k}$$

这是一个典型的二项式分布，当 T 趋于无穷大的时候（可以认为我们走的步伐特别短，要经过无穷多步才能遇到 k 个切点），它趋向于玻松分布，即极限情况是：

$$P(k) = e^{-\lambda} \lambda^k / k!$$

其中 $\lambda = T\rho$。上面得到的是在均匀的时间里得到 k 个切点的概率，下面由这个概率得到相邻两个切点需要经过 t 时间步，也就是线段长度的分布。设间隔时间 t 为随机变量，那么显然在这 T 个时间步内没有遇到切点的概率也就是遇到一个切点的间隔时间>T 的概率是：

$$P(t > T) = e^{-\lambda} = e^{-\rho T}$$

所以，随机变量 t 的概率密度函数就是：

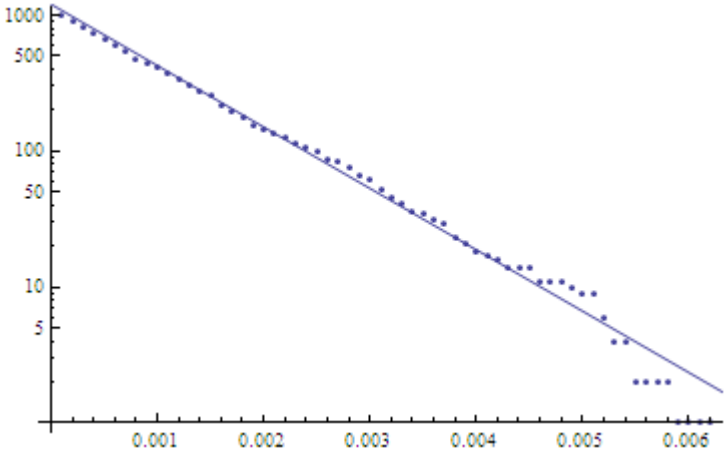
$$f(T) = F(t) = (1 - P(t > T))' = \rho e^{-\rho T}$$

因此，我们同样得到了负指数分布。

方法 4：计算机模拟

为了对这第一问中的分布有一个直观印象，我们可以做一个数值模拟，具体方法是，随机生

成[0,1]之间的 N 个随机数，它们构成均匀分布的切点，然后把它们排好序，求两两之间的差值，这就构成了线段长度的随机变量，求这个长度的分布。例如下面的图就是一个分布情况：



其中，纵轴我取了对数，横轴没取，所以指数函数所对应的分布就应该是直线。

2、模型 2

模型 2 相对简单，它是一个确定性的过程：
给定一个固定的比例数 k，从[0,1]之中切出两个区间：[0,k],(k,1)，然后再对[0,k]这个区间再切出比例 k 来，得到区间：[0,k^2],(k^2,k)，如此不断地重复下去……。
很显然最长的线段就是第一刀切完后右侧线段的长度，为(1-k)
第二长的线段为切第二刀后右侧线段的长度：k(1-k)
第二长的线段为切第二刀后右侧线段的长度：k^2(1-k)
……
所以，我们自然得到一个 Rank 曲线：

$$f(r) = k^r(1-k)$$

这是一个指数函数，把它画在纵坐标取对数的坐标轴上就得到了一条直线，这自然就有了书中图 5B 的那条曲线。

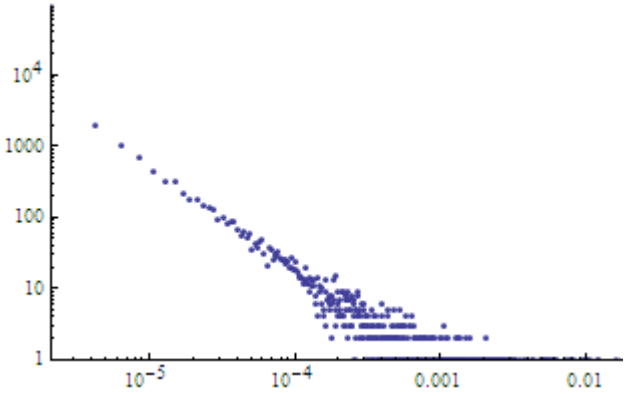
2、模型 3

这是第三种切线段模型，它与第一种很像，但是它能够得到对数正态分布。具体的思路是这样的：

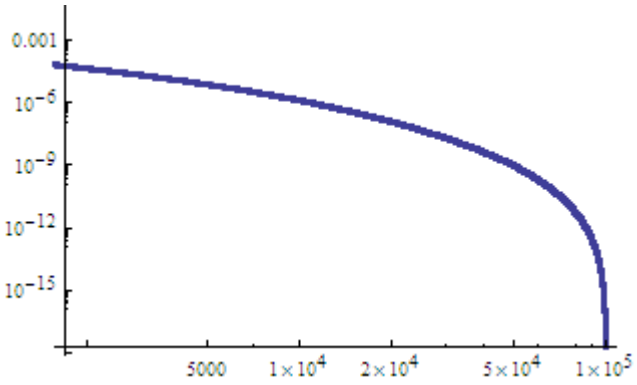
拿来一条单位长的线段，随机从中间砍一刀，得到两条线段，再从这两条里面随机拿一条，砍一刀，得到 3 条线段，再从这 3 条中随机选一条，砍一刀……，最后线段长度的分布是什么？

这个模型貌似跟第一个模型很相似，但是却是本质不同的，原因就在于在这个模型中选择线段切的时候是等概率选择的，也就是无论线段的长短都会等概率地被选中切一刀，这样短线段被选中的概率会大一些。

下面我们先给出这个问题的计算机模拟解，然后再进行数学分析。



线段长度的经验 PDF（双对数坐标轴）



线段长度的 Rank 曲线（双对数坐标轴）

虽然这个经验 PDF 曲线在双对数坐标下很像幂律分布，但是根据 Rank 曲线可知，这并不是幂律分布（如果是幂律的话，就应该是直线），而是对数正态分布（LogNormal）。

为什么会是对数正态分布呢？下面我们来证明一下。

假设我们现在已经砍过了 1 刀，下面我们再砍第 2 刀，假设这第 2 刀砍到了第 1 条线段上，这条线段的长度是β_0，它是一个[0, 1]之间的均匀分布的随机变量，随机这一刀切下去相当于是在[0,1]之间取了一个均匀分布的随机数β_1乘到了长度β_0上得到切点左边的那个小线段，之后切点右边的小线段长度就是$(1-\beta_1)*\beta_0$，也就是说在切了 2 刀后新添加的两个小线段的长度分别为：

$$x = \beta_1\beta_0$$
$$y = (1-\beta_1)\beta_0$$

然后，我们再切第 3 刀，假如这第 3 到刚好切到了刚才切过的小线段 x 上面，那么得到新的两个小线段长度就分别是：

$$x' = \beta_2\beta_1\beta_0$$
$$y' = (1-\beta_2)\beta_1\beta_0$$

以此类推我们可以得到 x”,x”’，假如我每一次选择都刚好选到了标号是 x 的小线段，那么新切得的小线段长度就可以写为：

$$x_n = \beta_0\beta_1\beta_2\beta_3\beta_4\ldots\beta_n$$

这相当于 n 个随机数的乘积。同理，如果没有选择 x，而是选择了 y，那么相乘的乘数

变成了(1-β)，但是它还是一个[0, 1]均匀分布的随机数。所以当切了无穷大刀数之后，每一条小线段的长度都可以看作是 n 个随机变量的乘积。

根据中心极限定理，经过简单变形，我们可以得到这 n 个随机变量的乘积在 n 趋近于无穷大的时候就会趋向于一个对数正态分布。原因是我们可以考察随机变量：

$$\log(x_n) = \sum_{i=1}^{n} \log(\beta_i)$$

我们已知 β_i 为一个[0,1]上面的均匀分布，所以 $\log(\beta_i)$ 的累积分布函数就是：

$$CDF(\log(\beta_i)) = P\{\log(\beta_i) < x < 0\} = P\{\beta_i < e^x < 1\} = \int_0^{e^x} 1dx = e^x$$

其中 x 为负数。这样 $\log(\beta_i)$ 的均值是：

$$m(\log(\beta_i)) = \int_{-\infty}^{0} xe^x dx = -1$$

方差是：

$$m(\log(\beta_i)) = \int_{-\infty}^{0} x^2 e^x dx = 2$$

所以 $\log(x_n)$满足均值为-1，方差为 2 的正态分布：

$$PDF(\log(x_n)) = e^{\frac{(x+1)^2}{4}} / 2\sqrt{2\pi}$$

因此，x_n 满足对数正态分布：

$$PDF(x_n) = e^{\frac{(1+\log x)^2}{8}} / 2\sqrt{2\pi}x$$

关于随机过程的进一步介绍，大家可参看王钺的PPT