**复杂网络与人类动力学中的常见分布律及数据拟合、参数估计**

**基本术语**

连续分布的概率密度函数 PDF：probability density function

离散分布的概率分布函数 PMF：probability mass function

连续分布的累积分布函数 CDF：cumulative distribution function，F(a)=P(x<a)

连续分布的互补累积分布函数 CCDF：complementary cumulative distribution function，F(a)=P(x>a)

方差：variance 标准差：standard deviation 均值：mean 期望：expectation

横坐标：abscissa 纵坐标：ordinate 坐标系：coordinate system

最小二乘回归：Ordinary least-square (OLS) regression 极大似然估计：Maximum likelihood estimation (MLE)

K-S检验：Kolmogorov-Smirnov test 拟合优度：Goodness-of-fit 显著性水平：Significance level

**常见的分布律**

* **正态/高斯分布Normal distribution / Gaussian distribution连续型**

正态分布是一种最重要最广泛的分布形式，和其它类型的分布（如泊松分布、二项分布等）有着密切关系。

The normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function or informally the bell curve.

PDF：其中为均值，为标准差。

当，时为标准正态分布standard normal distribution

PDF：

正态分布也经常写为：



正态分布的累积分布不常用，常用标准正态分布的累积分布。

* **对数正态分布Log-normal distribution连续型**

如果一个随机变量的对数服从正态分布，就称该随机变量服从对数正态分布。即：若X是服从正态分布的随机变量，则Y = exp(X)服从对数正态分布；若Y服从对数正态分布，则X = log(Y)服从正态分布。

PDF：

CDF：

where erfc is the complementary error function, and Φ is the cumulative distribution function of the standard normal distribution.

* **指数/负指数分布Exponential distribution / Negative exponential distribution连续型**

指数分布用于描述泊松过程的时间间隔，泊松过程中的事件以恒定速率连续且独立发生。是几何分布的连续类别。需要注意的是指数分布与指数分布族exponential families of distributions的区别，后者是一大类分布，包括指数分布、正态分布、二项分布、伽马分布、泊松分布等。

PDF：其中称为rate parameter，表示泊松过程的到达率。

CDF：

指数分布的PDF有时也会写成另外一种形式：

其中是的倒数，称为scale parameter。

指数分布的一个重要特征是无记忆性memoryless，即在等待时间已经超过s秒的前提下继续等待超过t秒的概率和一开始的时候等待时间超过t秒的概率相等。并且，指数分布和几何分布是唯一具有无记忆性的概率分布。

* **泊松分布Poisson distribution离散型**

泊松分布是离散型概率分布，用于描述在固定的时间或空间内，给定数量的事件发生的概率。事件以已知平均速率发生，且独立于时间。

PMF：

where e is the base of the natural logarithm (e = 2.71828...) and k! is the factorial of k.

* **幂律分布Power law distribution**
* 幂律分布的连续形式：

其中X是观测值，C是归一化常数，显然当时该式无意义，因此定义xmin来表示服从幂律的最小值。

当时有

* 幂律分布的离散形式：

同样计算归一化常数得到

其中是Hurwitz zeta function。

* 定义累积分布CDF：

连续形式：离散形式：

* 幂律分布广泛存在，其一个重要特征是标度不变性scale invariance，即



也就是说，具有相同标度指数的幂律分布实际上可以看作另一个幂律的scaled version，也正是这样的特征导致了的*f*(*x*)和*x*对数之间的线性关系，即在双对数坐标下呈现一条直线。但需要指出的是，不能仅仅凭借这样的直线就判定分布服从幂律，因为有其它分布在双对数坐标下也可能呈现直线形式。

* 幂律还具有普适性。如上所述，特有的标度指数可以区分不同的幂律，因此具有相同标度指数的幂律可视为一个普适类，于是就有了当年Vazquez等人将人类动力学分为幂指数分为1和1.5的两个离散的普适类。
* 此外，这种标度不变性本身就意味着一种自相似特征，即The ratio of the probabilities of occurrence of two sizes x1 and x2 depends only on their ratio x1/x2 and not on their absolute values. 并且越来越多的发现证实幂律与分形在复杂系统中确实有着密切联系。
* 幂律的判定有多种方法，一般都是判断其尾部是否服从幂律，因为双对数坐标下会更强调头部的作用。
* CDF：



显然，幂律分布的累积分布仍然服从幂律，只不过度指数会减一。

* 帕累托定律和齐普夫定律都属于幂律分布，前者描述的是个人收入X 不小于某个特定值x 的概率与x 的常数次幂存在简单的反比关系；后者描述的是每个单词出现的频率与它的名次的常数次幂存在的反比关系。实际上, 二者都是简单的幂律函数, Zipf 定律是幂律分布在有限值域上的离散形式, 而Pareto 定律是幂律分布的一种累积形式。
* 在有些文献中，幂律分布也会写成如下形式：

where L(x) is a slowly varying function defined by  for any finite t (typically, L(x) is a logarithm ln(x) or power of a logarithm such as  with n finite). In mathematical language, a function such as is said to be “regularly varying”.

* 在离散情况下，幂律分布的PDF往往还会用到Riemann-Zeta分布函数。
* **指数截断的幂律分布Power law with exponential cutoff**

带指数截断的幂律分布是幂律的一种重要变形，是一种幂律与指数混合的分布形式。其原理非常简单，在幂律项后面乘一个指数项即可。

PDF：或

指数截断中的指数衰减因子会在分布尾部超越幂律行为占据主导作用，这样的分布不是幂律的近似，而是在尾部之前的有限区域内有近似的标度行为。

This distribution is a common alternative to the asymptotic power-law distribution because it naturally captures finite-size effects. (关于finite-size effects详见EPJB 2004 38 205–209: Cut-offs and finite size effects in scale-free networks)

* **截断幂律Truncated power law**

这种分布是指数截断的幂律分布的另一种叫法，是一种幂律与指数混合的分布形式，在文献中，该分布也被视为double power law，且这样的幂律可由stretched exponential form拟合。详见下节。

PDF：

* **广延指数分布Stretched exponential distribution**

通常我们在双对数坐标下研究幂律，但双对数坐标有缺陷，并且实证中的幂律也很难在整个区间都表现为一条直线，于是有学者提出广延指数分布，作为整个区间上的一种替代性的分布。

PDF：

参数往往都小于等于1，显然当时即为指数分布，时为指数幂律混合分布，数值越小就越偏离指数而接近幂律，即在双对数坐标下表现出线性行为。

其它形式：PDF：

CDF：

类似的，当时为高斯分布，时为指数分布，时表现出重尾特征。

在上式两端取两次对数，可得

显然，若在双对数坐标下与的关系表现为一条直线，则为广延指数分布，且斜率即为参数的值。

如下图所示：



* **漂移幂律Shifted power law**

PDF：201112215921246.gif

漂移幂率也是一种综合了幂律与指数特征的混合分布形式，其中参数称为漂移量，可以控制分布在幂律()与指数()之间自由转换。这种分布在网络度分布理论中使用较多，在人类动力学中使用较少。

* 补充



Definition of the power-law distribution and several other common statistical distributions.

**数据拟合与参数估计**

* 我们已经知道，随机变量如果服从幂律分布的话会在双对数坐标下表现为一条直线。然而，直线的尾部可能由于统计误差而产生波动。解决尾部波动的一个方法是去掉尾部数据只拟合前半段，但这样做是一种舍本逐末的方法，会丢失其中包含的重要信息。另一种办法是进行装箱处理，但不论是线性装箱还是对数装箱也都属于粗粒化方法，仍然会产生统计噪声。因而常用最后一种方法：计算累积分布。使用累计分布不需要进行装箱操作，避免了确定装箱宽度的问题，并且对所有数据由较好的使用效果，不丢失任何信息。
* 数据拟合和参数估计有很多办法，包括计算累积分布、QQ图、Loglog图、Log-binning、OLS、MLE等等，目前的观点普遍认为双对数坐标下考察累积分布函数是正确的做法，幂律分布的指数应用极大似然方法估计，最后进行KS检验。基本步骤如下：

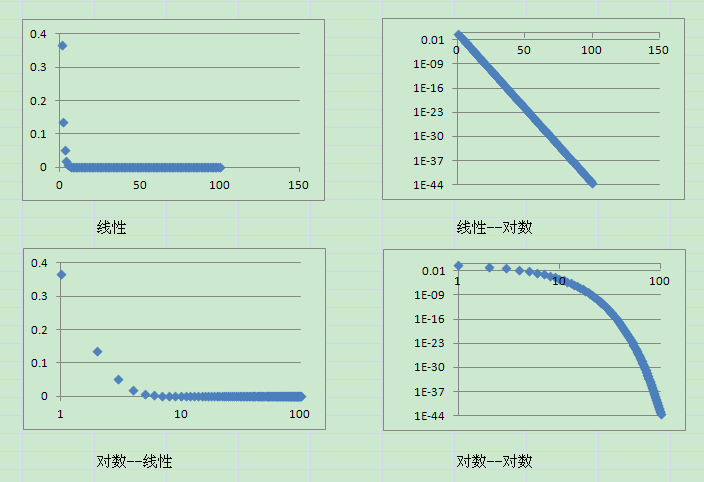
1. 构造Survivor distribution即the complementary cumulative distribution function，也就是统计物理中的累积分布，数学中的累积分布的补分布。
2. 原分布也一定要研究，因为累积分布并不是十全十美，可能在幂律的区间上让人难以判断，所以要结合原分布来判断是否是幂律，或者偏离幂律的根源。
3. 绘图，在线性、半对数、双对数坐标下都绘图，很多分布就可以判断出来了。如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
|  | linear-linear coordinates | log-linear coordinates | log-log coordinates |
| power law | 凸曲线（向下凸） | 凸曲线 | 直线 |
| Gaussian | 钟形曲线 | 反抛物线 | 明显下凹，急速衰减 |
| exponential | 凸曲线 | 直线 | 凹曲线 |

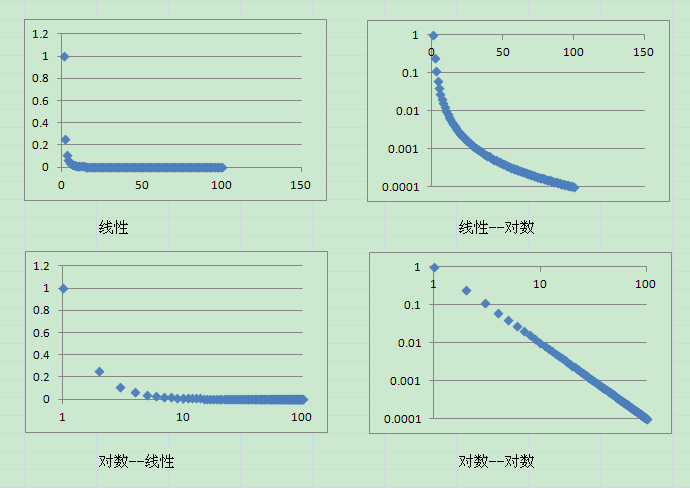
例如：若一个分布在双对数坐标下表现为向上凸起，说明其不是严格的幂律分布，在线性-对数坐标下观察，如果是直线，说明是指数分布，若向下凹，则是一种介于二者之间的分布，可能是Gamma分布，可能是广延指数分布。注意不要仅仅在双对数坐标系判断，因为双对数下表现为直线的分布不仅仅是幂律，可能是其它类型的分布！附图，通过图形判断分布类型。



文献Modelling collaboration networks based on nonlinear preferential attachment中对不同形态分布的描述。



指数分布在不同坐标下的图形



幂律分布在不同坐标下的图形

1. 对于幂律分布，因为当变量趋于0时会偏离幂律，故需要确定变量符合幂律区间的最小值。通常的方法要么在图形上凭视觉判断Xmin，要么是绘制幂指数与Xmin的散点图，这两种方法都比较主观，会受到噪声和波动的干扰。因而较准确的方法是进行KS估计，计算：



1. 对变量取对数后用最小二乘法进行参数拟合，得到参数的估计值。但是使用最小二乘法时隐含一个条件，即默认实际数据的对数与理论值的对数之间的偏差服从正态分布，这一假设并不合理，所以OLS法产生的估计是有偏估计。需要指出的是，所有基于图形的拟合都是最小二乘法。
2. 用极大似然法估计参数值。MLE方法通过极大化模型的似然函数来估计参数值，较为科学合理。对于幂律分布来说，常常按照下面的公式对其参数进行估计：

幂指数：其中xmin为x符合幂律部分的的最小值

标准差：

这两个公式都是连续形式下推导得出的，对于离散数据则复杂的多，一般也套用这样的公式。

离散整数数据也有下面的近似公式：

虽然说目前普遍认为最好的方法是累积分布+极大似然估计，但这两种方法也并非十全十美，累积分布的一个重要缺陷是在尾部往往会偏离幂律，故常用的方法是在拟合之前截掉尾部的数据。同样，头部的数据也常常会被去掉，因为现实中的幂律分布很难在整个区间内都很好的服从幂律。

1. 上一步进行参数估计的时候是假设数据服从某种分布的，而这种假设或者猜想未必正确，因此需要通过KS检验进行判定。具体步骤如下：

①按照下面的公式计算实际值和理论值的累积分布之间的最大距离

，得到一个D值；

②令N=数据量，即整个样本中数据的个数，在下表中查找对应的样本量。例如：如果N=1354，那么取表中的N=1000，当然可以选择更高的2000，但实际中样本量如果不是很接近临界值如N=1900，选择近似值即可；



③然后在下表中查找最接近的分位数，如N=1000对应的分位数0.0186，而刚才计算得到的D=0.0117，小于该分位数，这就意味着在该显著性水平下，有超过10%的几率该分布服从幂律，因此拒绝幂律假设的证据不足，可以认定该分布服从幂律。

**重要参考文献**

* Ming-Sheng Shang, Linyuan Lv, Yi-Cheng Zhang and Tao Zhou. Empirical analysis of web-based user-object bipartite networks. EPL, 90 (2010) 48006. （二部图，广延指数分布）
* TAO ZHOU, BING-HONG WANG, et al. MODELLING COLLABORATION NETWORKS BASED ON NONLINEAR PREFERENTIAL ATTACHMENT. International Journal of Modern Physics C, 18( 2) (2007) 297-314. （网络的一些基本概念，truncated power law，广延指数分布）
* Jean LAHERRERE and Didier SORNETTE. STRETCHED EXPONENTIAL DISTRIBUTIONS IN NATURE AND ECONOMY: FAT TAILS WITH CHARACTERISTIC SCALES.（幂律--分形，广延指数分布的性质）
* M.L. Goldstein, S.A. Morris, and G.G. Yen. Problems with fitting to the power-law distribution. Eur. Phys. J. B 41, 255–258 (2004). （KS检验实例，分位数表）
* Heiko Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. Eur. Phys. J. B 58, 167-173 (2007). （极大似然估计）
* Fitting empirical distributions to theoretical models. （分布简介与Matlab代码R代码）
* Andreas Klaus, Shan Yu, Dietmar Plenz. Statistical Analyses Support Power Law Distributions Found in Neuronal Avalanches. May 2011, 6(5): e19779. （分布简介，OLS，MLE，KS简介）
* DIDIER SORNETTE. Probability Distributions in Complex Systems. （PDF、CDF等概念，各种分布的概念，数据拟合的整体步骤）
* M. E. J. Newman. The structure of scientific collaboration networks. 404–409, PNAS, January 16, 2001, 98(2). （指数截断的幂律分布）
* ETHAN P. WHITE, BRIAN J. ENQUIST, AND JESSICA L. GREEN. ON ESTIMATING THE EXPONENT OF POWER-LAW FREQUENCY DISTRIBUTIONS. Ecology, 89(4), 2008, 905–912. （幂律分布的简介、变形、线性装箱、对数装箱、CDF、MLE、拟合方法及效果对比）
* Newman MEJ (2005) Power laws, pareto distributions and Zipf’s law. Contemporary Physics 46: 323–351. （幂律分布简介，拟合方法，参数估计，幂律的实例，变形，数学性质，离散形式）
* Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51: 661–703. （幂律定义，分析步骤，LS，ML，各种参数估计方法的对比，如何确定符合幂律的最小值Xmin，KS检验的严格步骤）
* M. Boguna, R. Pastor-Satorras, and A. Vespignani. Cut-offs and finite size effects in scale-free networks. Eur. Phys. J. B 38, 205–209 (2004). （文献中常提到的finite size effects）