正态分布的前世今生(下)

靳志辉

六、开疆拓土，正态分布的进一步发展

19 世纪初，随着拉普拉斯中心极限定理的建立与高斯正态误差理论的问世，正态分布开始崭露头角，逐步在近代概率论和数理统计学中大放异彩。在概率论中，由于拉普拉斯的推动，中心极限定理发展成为现代概率论的一块基石；而在数理统计学中，在高斯的大力提倡之下，正态分布开始逐步畅行于天下。

1 论剑中心极限定理

先来说说正态分布在概率论中的地位，这个主要是由于中心极限定理的影响。1776 年，拉普拉斯开始考虑一个天文学中彗星轨道的倾角的计算问题，最终的问题涉及独立随机变量求和的概率计算，也就是计算如下的概率值

$$S_n = X_1 + X_2 + \cdots + X_n$$
$$P(a < S_n < b) = ?$$

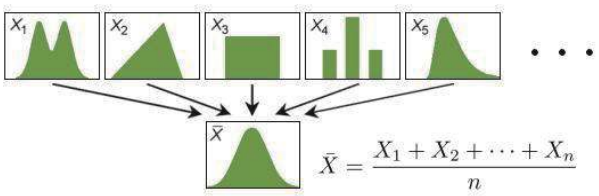
在这个问题的处理上，拉普拉斯充分展示了其深厚的数学分析功底和高超的概率计算技巧，他首次引入了用特征函数（也就是对概率密度函数做傅立叶变换）来处理概率分布的神妙方法，而这一方法经过几代概率学家的发展，在现代概率论里面占有极其重要的位置。基于这一分析方法，拉普拉斯通过近似计算，在他 1812 年发表的名著《分析概率论》中给出了中心极限定理的一般描述：

定理 0.6.1 (拉普拉斯，1812) $e_i\,(i = 1, \cdots n)$ 为独立同分布的测量误差，具有均值 μ 和方差 σ^2，如果 $\lambda_1, \cdots, \lambda_2$ 为常数，$\alpha > 0$, 则有

$$P\left(\left|\sum_{i=1}^{n}\lambda_i(e_i-\mu)\right| \le a\sqrt{\sum_{i=1}^{n}\lambda_i^2}\right) \approx \frac{2}{\sqrt{2\pi}\sigma}\int_0^a e^{-\frac{x^2}{2\sigma^2}}dx.$$

这已经是比棣莫弗 - 拉普拉斯中心极限定理更加深刻的一个结论了，理科专业的本科生学习《概率论与数理统计》这门课程的时候，通常学习的中心极限定理的一般形式如下：

定理 0.6.2 (林德伯格 - 列维中心极限定理) 设 $X_1, \cdots, X_n$ 独立同分布，且具有有限的均值 μ 和方差 σ^2，则在 $n \to \infty$ 时，有



中心极限定理

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \to N(0,1).$$

多么奇妙的性质，随意的一个概率分布中生成的随机变量，在序列和 (或者等价的求算术平均) 的操作之下，表现出如此一致的行为，统一地规约到正态分布。

概率学家们进一步的研究结果更加令人惊讶，序列求和最终要导出正态分布的条件并不需要这么苛刻，即便 $X_1, \cdots, X_n$ 并不独立，也不具有相同的概率分布形式，很多时候他们求和的最终归宿仍然是正态分布。一切的纷繁芜杂都在神秘的正态曲线下被消解，这不禁令人浮想联翩。中心极限定理恐怕是概率论中最具有宗教神秘色彩的定理，如果有一位牧师拿着一本圣经向我证明上帝的存在，我是丝毫不会买账；可是如果他向我展示中心极限定理并且声称那是神迹，我可能会有点犹豫，从而乐意倾听他的布道。如果我能坐着时光机穿越到一个原始部落中，我也一定带上中心极限定理，并劝说部落的酋长把正态分布作为他们的图腾。

中心极限定理虽然表述形式简洁，但是严格证明它却非常困难。中心极限定理就像一张大蜘蛛网，棣莫弗和拉普拉斯编织了它的雏形，可是这张网上漏洞太多，一个多世纪来，数学家们就像蜘蛛一样前赴后继，努力想把所有的漏洞都补上。在 19 世纪，泊松（Siméon Denis Poisson, 1781-1840）、狄利克雷（Gustav Lejeune Dirichlet, 1805-1859）、柯西（Augustin-Louis Cauchy, 1789-1857）、贝塞尔（Friedrich Bessel, 1784-1846）这些大蜘蛛都曾经试图把这张网的漏洞补上。从现代概率论的角度来看，整个 19 世纪的经典概率理论并没有能输出一个一般意义下的严格证明。而最终把漏洞补上的是来自俄罗斯的几位蜘蛛侠：切比雪夫（Pafnuty Chebyshev, 1821-1894）、马尔可夫（Andrey Andreyevich Markov, 1856-1922）和李雅普诺夫（Aleksandr Mikhailovich Lyapunov, 1857-1918）。俄罗斯是一个具有优秀数学传统的民族，产生过几位顶尖的数学家，在现代概率论的发展中，俄罗斯的圣彼得堡学派可以算是顶了大半边天，而切比雪夫正是圣彼得堡数学学派的奠基人和领袖。给中心极限定理补漏的方案雏形是从切比雪夫 1887 年的工作开始的，切比雪夫提出了一个基于矩法的证明，矩法是概率分析中比较传统的方法，使用的数学工具比较基础，不过切比雪夫这个证明也还存在一些漏洞。马尔可夫和李雅普诺夫都是切比雪夫的学生，两人在中心极限定理的严格证明上展开了竞赛。马尔可夫在概率论里面可算是大名鼎鼎，马尔可夫链是应用最为广泛的概率模型之一。马尔可夫和他的老师切比雪夫一样，他们在数学中的研究风格都偏向于使用初等、简单易懂的数学工具来证明复杂艰深的定理。马尔可夫沿着老师的基于矩法的思路在蜘蛛网上辛勤编织，在证明上做了很多的修补，但洞还是补得不够严实。李雅普诺夫不像马尔可夫那样深受老师的影响，他沿着拉普拉斯当年提出的基于特征函数的思路，于 1901 年给出了一个补洞的方法，切比雪夫对这个方法大加赞赏，李雅普诺夫的证明被认为是第一个在一般条件下的严格证明；而马尔可夫也不甘示弱，在 1913 年基于矩法也把洞给补严实了。

20 世纪初期到中期，中心极限定理的研究几乎吸引了所有的概率学家，这个定理俨然成为了概率论的明珠，成为了各大概率论武林高手华山论剑的场所。不知道大家对中心极限定理中的“中心”一词如何理解，许多人都认为“中心”这个词描述的是这个定理的行为：以正态分布为中心。这个解释看起来确实合情合理，不过并不符合该定理被冠名的历史。事实上，20 世纪初概率学家大都称呼该定理为极限定理（Limit Theorem），由于该定理在概率论中处于如此重要的中心位置，如此之多的概率学武林高手为它魂牵梦绕，于是数学家波利亚于 1920 年在该定理前面冠以“中心”一词，由此后续人们都称之为中心极限定理。

数学家们总是极其严谨苛刻的，给定的一个条件下严格证明了中心极限定理，数学家们就开始探寻中心极限定理成


切比雪夫（1821-1894）


马尔可夫（1856-1922）


李雅普诺夫（1857-1918）


费勒（1906-1970）


列维（1886-1971）

华山论剑

立的各种条件，询问这个条件是否为充分必要条件，并且进一步追问序列和在该条件下以什么样的速度收敛到正态分布。1922 年林德伯格（Jarl Waldemar Lindeberg, 1876-1932）基于一个比较宽泛且容易满足的条件，为中心极限定理提出了一个很容易理解的初等证明，这个条件我们现在称之为林德伯格条件。然后概率学家费勒（William Feller, 1906-1970）和列维（Paul Pierre Lévy, 1886-1971）就开始追问：林德伯格条件是充分必要的吗？基于林德伯格的工作，费勒和列维都于 1935 年独立地得到了中心极限定理成立的充分必要条件，这个条件可以用直观的非数学语言描述如下：

定理 0.6.3（中心极限定理充要条件）假设独立随机变量序列 X_i 的中值为 0，要使序列和 $S=\sum_{i=1}^{n} X_i$ 的分布密度函数逼近正态分布，以下条件是充分必要的：

* 如果 X_i 相对于序列和 S 的散布（可以理解为标准差）是不可忽略的，则 X_i 的分布必须接近正态分布；

* 对于所有可忽略的 X_i，取绝对值最大的那一项，这个绝对值相对于序列和也是可忽略的。

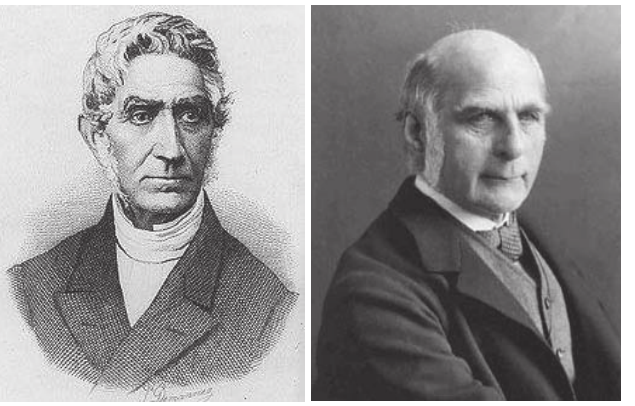
事实上这个充分必要条件发现的优先权，费勒和列维之间还着实出现了一些争论，当然他们俩都是独立的几乎在同一时间解决了这个问题。在列维证明这个充分必要条件的过程中，列维发现了正态分布的一个有趣的性质。我们在数理统计中都学过，如果两个独立随机变量 X, Y 具有正态分布，则 $S=X+Y$ 也具有正态分布；奇妙的是这个定理的逆定理也成立：

定理 0.6.4（正态分布的血统）如果 X, Y 是独立的随机变量，且 $S=X+Y$ 是正态分布，那么 X, Y 也是正态分布。

正态分布真是很奇妙，就像蚯蚓一样具有再生的性质，你把它一刀切两断，它生成两个正态分布；或者说正态分布具有极其纯正的优良血统，正态分布的组成成分中只能包含正态分布，而不可能含有其它杂质。

一流的数学家都是接近上帝的人，善于猜测上帝的意图。1928 年列维就猜到了这个定理，并在 1935 年使用这个定理对中心极限定理的充分必要条件作了证明。有意思的是列维却无法证明正态分布的这个看上去极其简单的再生性质，所以他的证明多少让人觉得有些瑕疵。不过列维的救星很快就降临了，1936 年概率学家克拉美（Harald Cramér, 1893-1985）证明列维的猜想完全正确。

中心极限定理成为了现代概率论中首屈一指的定理，事实上中心极限定理在现代概率论里面已经不是指一个定理，而是指一系列相关的定理。统计学家们也基于该定理不断地完善拉普拉斯提出的元误差理论，并据此解释为何世界上正态分布如此常见。而中心极限定理同时成为了现代统计学中大样本理论的基础。



凯特勒（1796-1874）和高尔顿（1822-1911）

2 进军近代统计学

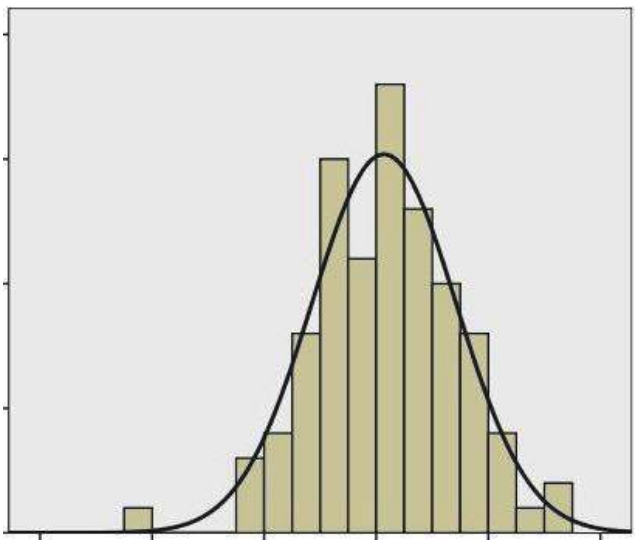
花开两朵，各表一枝。上面说了正态分布在概率论中的发展，现在来看看正态分布在数理统计学中发展的故事。这个故事的领衔主演是凯特勒（Adolphe Quetelet, 1796-1874）和高尔顿（Francis Galton, 1822-1911）。

由于高斯的工作，正态分布在误差分析中迅速确定了自己的地位。有了这么好的工具，我们可能拍脑袋就认为，正态分布很快就被人们用来分析其它的数据，然而事实却出乎我们的意料，正态分布进入社会领域和自然科学领域，可是经过一番周折的。

首先我要告诉大家一个事实：误差分析和统计学是风马牛不相及的两个学科，当然，这个事实存在的时限截止到 19 世纪初。统计学的产生最初是与“编制国情报告”有关，主要服务于政府部门。统计学面对的是统计数据，是对多个不同对象的测量；而误差分析研究的是观测数据，是对同一个对象的多次测量。观测数据和统计数据在当时被认为是两种不同行为获取得到的数据，适用于观测数据的规律未必适用于统计数据。19 世纪的数据统计分析处于一个很落后的状态，和概率论没有多少结合。概率论的产生主要和赌博相关，发展过程中与误差分析紧密联系，而与当时的统计学交集非常小。将统计学与概率论真正结合起来推动数理统计学发展的便是我们的统计学巨星凯特勒。

凯特勒这名字或许不如其它数学家那么响亮，估计很多人不熟悉，所以有必要介绍一下。凯特勒是比利时人，数学博士毕业，年轻的时候曾追随拉普拉斯学习过概率论。此人学识渊博，涉猎广泛，脑袋上的桂冠包括统计学家、数学家、天文学家、社会学家、国际统计会议之父、近代统计学之父、数理统计学派创始人。凯特勒的最大贡献就是将法国的古典概率理论引入统计学，用纯数学的方法对社会现象进行研究。

1831 年，凯特勒参与主持新建比利时统计总局的工作。



用正态分布拟合数据

他开始从事有关人口问题的统计学研究。在这种研究中，凯特勒发现，以往被人们认为杂乱无章的、偶然性占统治地位的社会现象，如同自然现象一样也具有一定的规律性。凯特勒搜集了大量关于人体生理测量的数据，如体重、身高与胸围等，并使用概率统计方法来对数据进行数据分析。但是当时的统计分析方法遭到了社会学家的质疑，社会学家们的反对意见主要在于：社会问题与科学实验不同，其数据一般由观察得到，无法控制且经常不了解其异质因素，这样数据的同质性连带其分析结果往往就有了问题，于是社会统计工作者就面临一个如何判断数据同质性的问题。凯特勒大胆地提出：**把一批数据是否能很好地拟合正态分布，作为判断该批数据是否同质的标准**。

凯特勒提出了一个使用正态曲线拟合数据的方法，并广泛地使用正态分布去拟合各种类型的数据。由此，凯特勒为正态分布的应用拓展了广阔的舞台。正态分布如同一把屠龙刀，在凯特勒的带领下，学者们挥舞着这把宝刀在各个领域披荆斩棘，攻陷了人口、领土、政治、农业、工业、商业、道德等社会领域，并进一步攻占天文学、数学、物理学、生物学、社会统计学及气象学等自然科学领域。

正态分布的下一个推动力来自生物学家高尔顿，当正态分布与生物学联姻时，近代统计学迎来了一次大发展。高尔顿是生物统计学派的奠基人，他的表哥达尔文的巨著《物种起源》问世以后，触动他用统计方法研究遗传进化问题。受凯特勒的启发，他对正态分布怀有浓厚的兴趣，开始使用正态分布去拟合人的身高、胸围以至考试成绩等各类数据，发现正态分布拟合得非常好。他因此相信正态曲线是适用于无数情况的一般法则。

然而，对高尔顿而言，这个无处不在的正态性给他带来一些困惑。他考察了亲子两代的身高数据，发现遵从同一的正态分布，遗传作为一个显著因素是如何发挥作用的？1877年，高尔顿设计了一个叫高尔顿钉板（quincunx, 或者 Galton board）的装置，模拟正态分布的性质，用于解释遗传现象。

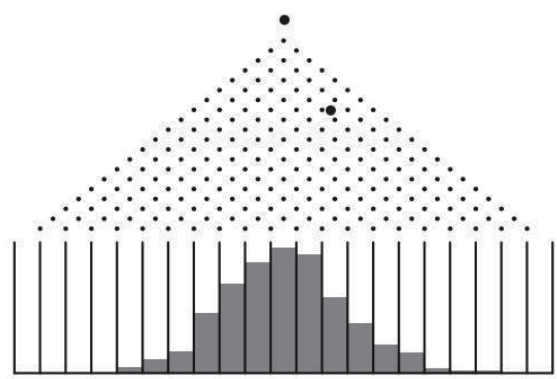
如下图中每一点表示钉在板上的一颗钉子，它们彼此的距离均相等。当小圆球向下降落过程中，碰到钉子后皆以1/2 的概率向左或向右滚下。如果有 n 排钉子，则各槽内最终球的个数服从二项分布 $B(n, 1/2)$，当 n 较大的时候，接近正态分布。

设想在此装置的中间某个地方 AB 设一个挡板把小球截住，小球将在 AB 处聚成正态曲线形状，如果挡板上有许多阀门，打开一些阀门，则在底部形成多个大小不一的正态分布，而最终的大正态分布正是这些小正态分布的混合。

高尔顿利用这个装置创造性地把正态分布的性质用于解释遗传现象。他解释说身高受到显著因素和其它较小因素的影响，每个因素的影响可以表达为一个正态分布。遗传作为一个显著因素，类似图中底部大小不一的正态分布中的比较大的正态分布，而多个大小不一正态分布累加之后其结果仍然得到一个正态分布。

网络上有人开发了一些好玩的程序用于模拟高尔顿钉板，我们可以动态地观察这些小球自上而下滚动的时候，是如何形成正态分布的。一个使用 Java 开发的很漂亮的动态模拟可以在如下网页中观察到：http://www.math.psu.edu/dlittle/java/probability/plinko/index.html。

高尔顿在研究身高的遗传效应的时候，同时发现一个奇特的现象：高个子父母的子女，其身高有低于其父母身高的趋势，而矮个子父母的子女，其身高有高于其父母的趋势，即有“回归”到普通人平均身高去的趋势，这也是“回归”一词最早的含义。高尔顿用二维正态分布去拟合父代和子代身高的数据，同时引进了回归直线、相关系数的概念，从而开创了回归分析这门技术。



高尔顿钉板