正态分布的前世今生(上)

靳志辉

神说，要有正态分布，就有了正态分布。
神看正态分布是好的，就让随机误差服从了正态分布。
创世纪——数理统计

一、正态分布——熟悉的陌生人

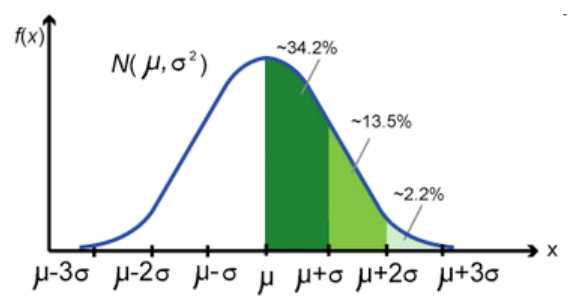
学过基础统计学的同学大都对正态分布非常熟悉。这个钟形的分布曲线不但形状优雅，其密度函数写成数学表达式

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

也非常具有数学的美感。其标准化后的概率密度函数

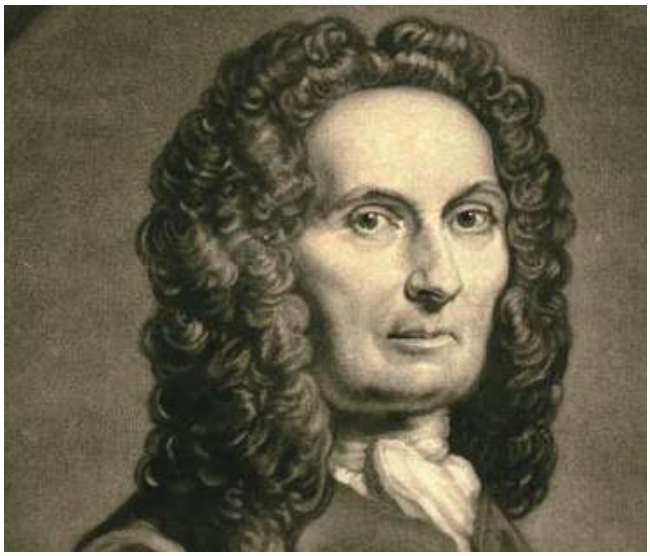
$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

更加的简洁漂亮，两个最重要的数学常量 π, e 都出现在这公式之中。在我个人的审美之中，它也属于 top-N 的最美丽的数学公式之一，如果有人问我数理统计领域哪个公式最能让人感觉到上帝的存在，那我一定投正态分布的票。因为这个分布戴着神秘的面纱，在自然界中无处不在，让你在纷繁芜杂的数据背后看到隐隐的秩序。



正态分布曲线

正态分布又通常被称为高斯分布，在科学领域，冠名权那是一个很高的荣誉。2002 年以前去过德国的兄弟们还会发现，德国 1991 年至 2001 年间发行的的一款 10 马克的纸币上印着高斯（Carl Friedrich Gauss, 1777-1855）的头像和正态密度曲线，而 1977 年东德发行的 20 马克的可流通纪



棣莫弗（1667-1754）

拉普拉斯（1749-1827）

念钢镚上，也印着正态分布曲线和高斯的名字。正态分布被冠名高斯分布，我们也容易认为是高斯发现了正态分布，其实不然，不过高斯对于正态分布的历史地位的确立是起到了决定性的作用。

正态曲线虽然看上去很美，却不是一拍脑袋就能想到的。我们在本科学习数理统计的时候，课本一上来介绍正态分布就给出密度分布函数，却从来不说明这个分布函数是通过什么原理推导出来的。所以我一直搞不明白数学家当年是怎么找到这个概率分布曲线的，又是怎么发现随机误差服从这个奇妙的分布的。我们在实践中大量地使用正态分布，却对这个分布的来龙去脉知之甚少，正态分布真是让人感觉既熟悉又陌生。直到我读研究生的时候，我的导师给我介绍了陈希孺院士的《数理统计学简史》这本书，看了之后才了解到正态分布曲线从发现到被人们重视进而广泛应用，也是经过了几百年的历史。

正态分布的这段历史是很精彩的，我们通过讲一系列的故事来揭开她的神秘面纱。

二、邂逅——正态曲线的首次发现

第一个故事和概率论的发展密切相关，主角是棣莫弗((Abraham de Moivre, 1667-1754）和拉普拉斯（Pierre-Simon Laplace, 1749-1827）。拉普拉斯是个大科学家，被称为法国的牛顿；棣莫弗名气可能不算很大，不过大家应该都熟悉这个名字，因为我们在高中数学学复数的时候都学过棣莫弗公式 $(\cos\theta+i\sin\theta)^n=\cos(n\theta)+i\sin(n\theta)$。

古典概率论发源于赌博，惠更斯（Christiaan Huygens, 1629-1695）、帕斯卡 (Blaise Pascal, 1623-1662)、费马（Pierre de Fermat, 1601-1665）、雅可比•贝努利（Jacob Bernoulli, 1654-1705）都是古典概率的奠基人，他们那会儿研究的概率问题大都来自赌桌上，最早的概率论问题是赌徒梅累在1654年向帕斯卡提出的如何分赌金的问题。统计学中的总体均值之所以被称为期望（Expectation），就是源自惠更斯、帕斯卡这些人研究平均情况下一个赌徒在赌桌上可以期望自己赢得多少钱。

有一天一个哥们，也许是个赌徒，向棣莫弗提了一个和赌博相关的问题：A、B两人在赌场里赌博，A、B各自的获胜概率是 $p, q=1-p$，赌 n 局，两人约定：若A赢的局数 $X>np$，则A付给赌场 $X-np$ 元，若 $X<np$，则B付给赌场 $np-X$ 元。问赌场挣钱的期望值是多少？

问题并不复杂，本质上是一个二项分布，若 np 为整数，棣莫弗求出最后的理论结果是

$$2npqb(n,p,np).$$

其中

$$b(n,p,i)=\binom{n}{i}p^iq^{n-i}$$

是常见的二项概率。但是对具体的 n，因为其中的二项公式中有组合数，要把这个理论结果实际计算出数值结果可不是件容易的事，这就驱动棣莫弗寻找近似计算的方法。

与此相关联的另一个问题，是遵从二项分布的随机变量 $X\sim B(n,p)$，求 X 落在二项分布中心点一定范围的概率 $P_d=P(|X-np|\le d)$。

对于 $p = 1/2$ 的情形，棣莫弗做了一些计算并得到了一些近似结果，但是还不够漂亮，幸运的是棣莫弗和斯特林（James Stirling, 1692-1770）处在同一个时代，而且二人之间有联系，斯特林公式是在数学分析中必学的一个重要公式：

$$n!\approx\sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

事实上斯特林公式的形式其实是棣莫弗最先发现的，但是斯特林改进了这个公式，改进的结果为棣莫弗所用。1733 年，棣莫弗很快利用斯特林公式进行计算并取得了重要的进展。考虑 n 是偶数的情形，二项概率为

$$b(n,\frac{1}{2},i)=\binom{n}{i}\left(\frac{1}{2}\right)^n.$$

以下把 $b(n,\frac{1}{2},i)$ 简记为 $b(i)$，通过斯特林公式做一些简单的计算容易得到

$$b\left(\frac{n}{2}\right)\approx\sqrt{\frac{2}{\pi n}},\qquad \frac{b(\frac{n}{2}+d)}{b(\frac{n}{2})}\approx e^{-\frac{2d^2}{n}}.$$

于是有

$$b\left(\frac{n}{2}+d\right)\approx\frac{2}{\sqrt{2\pi n}}e^{-\frac{2d^2}{n}}.$$

使用上式的结果，并在二项概率累加求和的过程中近似地使用定积分代替求和，很容易就能得到

$$\begin{aligned}P\left(\left|\frac{X}{n}-\frac{1}{2}\right|\le\frac{c}{\sqrt{n}}\right)&=\sum_{-c\sqrt{n}\le i\le c\sqrt{n}}b\left(\frac{n}{2}+i\right)\\&\approx\sum_{-c\sqrt{n}\le i\le c\sqrt{n}}\frac{2}{\sqrt{2\pi n}}e^{-\frac{2i^2}{n}}\\&=\sum_{-2c\sqrt{n}\le\frac{2i}{\sqrt{n}}\le 2c}\frac{2}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{2i}{\sqrt{n}}\right)^2}\frac{2}{\sqrt{n}}\\&\approx\int_{-2c}^{2c}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\end{aligned}\qquad(1)$$

看，正态分布的密度函数的形式在积分公式中出现了！这也就是我们在数理统计课本上学到的一个重要结论：二项分布的极限分布是正态分布。

以上只是讨论了 $p = 1/2$ 的情形，棣莫弗也对 $p \neq 1/2$ 做了一些计算，后来拉普拉斯对 $p \neq 1/2$ 的情况做了更多的分析，并把二项分布的正态近似推广到了任意 p 的情况。这是第一次正态密度函数被数学家刻画出来，而且是以二项分布的极限分布的形式被推导出来的。熟悉基础概率统计的同学们都知道这个结果其实叫棣莫弗 - 拉普拉斯中心极限定理。

[棣莫弗 - 拉普拉斯中心极限定理] 设随机变量 $X_n(n=1,2,\cdots\cdots)$ 服从参数为 n 和 p 的二项分布，则对任意的 x，恒有

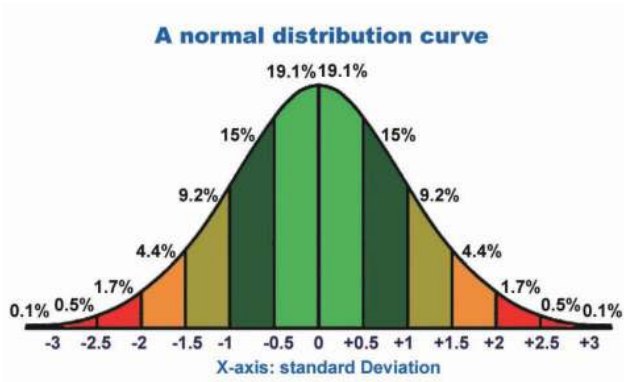
陈希孺编著的《数理统计学简史》

$$\lim_{n\to\infty}P\left(\frac{X_n-np}{\sqrt{np(1-p)}}\le x\right)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}dt.$$

我们在大学学习数理统计的时候，学习的过程都是先学习正态分布，然后才学习中心极限定理。而学习到正态分布的时候，直接就描述了其概率密度的数学形式，虽然数学上很漂亮，但是容易困惑数学家们是如何凭空就找到这个分布的。读了陈希孺的《数理统计学简史》之后，我才明白正态分布的密度形式首次发现是在棣莫弗 - 拉普拉斯的中心极限定理中。数学家研究数学问题的进程很少是按照我们数学课本的安排顺序推进的，现代的数学课本都是按照数学内在的逻辑进行组织编排的，虽然逻辑结构上严谨优美，却把数学问题研究的历史痕迹抹得一干二净。DNA双螺旋结构的发现者之一詹姆斯 • 沃森（James D. Watson, 1928-）在他的名著《DNA双螺旋》序言中说：“Science seldom proceeds in the straightforward logical manner imagined by outsiders.（科学的发现很少会像门外汉所想象的那样按照直接了当合乎逻辑的方式进行。）”

棣莫弗给出他的发现后 40 年（大约是 1770 年），拉普拉斯建立了中心极限定理较一般的形式，中心极限定理随后又被其他数学家们推广到了其他任意分布的情形，而不限于二项分布。后续的统计学家发现，一系列的重要统计量，在样本量 N 趋于无穷的时候，其极限分布都有正态的形式，这构成了数理统计学中大样本理论的基础。

棣莫弗在二项分布的计算中瞥见了正态曲线的模样，不过他并没有能展现这个曲线的美妙之处。棣莫弗的这个工作



最小二乘法的一个例子

当时并没有引起人们足够的重视，原因在于棣莫弗不是个统计学家，从未从统计学的角度去考虑其工作的意义。正态分布（当时也没有被命名为正态分布）在当时也只是以极限分布的形式出现，并没有在统计学，尤其是误差分析中发挥作用。这也就是正态分布最终没有被冠名棣莫弗分布的重要原因。那高斯做了啥了不起的工作导致统计学家把正态分布的这顶桂冠戴在了他的头上呢？这先得从最小二乘法的发展说起。

三、最小二乘法——数据分析的瑞士军刀

第二个故事的主角是欧拉（Leonhard Euler, 1707-1783）、拉普拉斯、勒让德（Adrien-Marie Legendre, 1752-1833）和高斯，故事发生的时间是十八世纪中到十九世纪初。十七、十八世纪是科学发展的黄金年代，微积分的发展和牛顿万有引力定律的建立，直接地推动了天文学和测地学的迅猛发展。当时的大科学家们都在考虑许多天文学上的问题。几个典型的问题如下：

* 土星和木星是太阳系中的大行星，由于相互吸引对各自的运动轨道产生了影响，许多大数学家，包括欧拉和拉普拉斯都基于长期积累的天文观测数据计算土星和木星的运行轨道。
* 勒让德承担了一个政府给的重要任务，测量通过巴黎的子午线的长度。
* 海上航行经纬度的定位。主要是通过对恒星和月面上的一些定点的观测来确定经纬度。

这些天文学和测地学的问题，无不涉及到数据的多次测量、分析与计算；十七、十八世纪的天文观测，也积累了大量的数据需要进行分析和计算。很多年以前，学者们就已经经验性地认为，对于有误差的测量数据，多次测量取算术平均是比较好的处理方法。虽然缺乏理论上的论证，也不断

勒让德（1752-1833）

地受到一些人的质疑，取算术平均作为一种异常直观的方式，已经被使用了千百年，在多年积累的数据的处理经验中也得到相当程度的验证，被认为是一种良好的数据处理方法。

以上涉及的问题，我们直接关心的目标量往往无法直接观测，但是一些相关的量是可以观测到的，而通过建立数学模型，最终可以解出我们关心的量。这些问题都可以用如下数学模型描述：我们想估计的量是 $\beta_0, \cdots, \beta_p$，另有若干个可以测量的量 $x_1, \cdots, x_p, y$，这些量之间有线性关系

$$y = \beta_0+\beta_1 x_1+ \cdots +\beta_p x_p,$$

如何通过多组观测数据求解出参数 $\beta_0, \cdots, \beta_p$ 呢？欧拉和拉普拉斯采用的都是求解线性方程组。

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{p1} \\ y_2 = \beta_0 + \beta_1 x_{12} + \cdots + \beta_p x_{p2} \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{1n} + \cdots + \beta_p x_{pn} \end{cases} \tag{2}$$

但是面临的一个问题是，有 n 组观测数据，p+1 个变量，如果 $n>p$+1，则得到的线性矛盾方程组无法直接求解。所以欧拉和拉普拉斯采用的方法都是通过对数据一定的观察，把 n 个线性方程分为 p+1 组，然后把每个组内的方程线性求和后归并为一个方程，从而就把 n 个方程的方程组化为 p+1 个方程的方程组，进一步解方程求解参数。这些方法初看有一些道理，但是都过于经验化，无法形成统一处理这一类问题的通用解决框架。

以上求解线性矛盾方程的问题在现在的本科生看来都

不困难，这就是统计学中的线性回归问题，直接用最小二乘法就解决了。可是即便如欧拉、拉普拉斯这些数学大牛，当时也未能对这些问题提出有效的解决方案。可见在科学研究中，要想在观念上有所突破并不容易。有效的最小二乘法是勒让德在 1805 年发表的，基本思想就是认为测量中有误差，所以所有方程的累积误差为

$$\text{累积误差}=\sum(\text{观测值}-\text{理论值})^2$$

我们求解出导致累积误差最小的参数：

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} e_i^2 \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) \right]^2 \end{aligned} \tag{3}$$

勒让德在论文中对最小二乘法的优良性做了几点说明：

* 最小二乘法使得误差平方和最小，并在各个方程的误差之间建立了一种平衡，从而防止某一个极端误差取得支配地位。
* 计算中只要求偏导后求解线性方程组，计算过程明确便捷。
* 最小二乘法可以导出算术平均值作为估计值。

对于最后一点，推理如下：假设真值为 $\theta, x_1, \cdots, x_n$ 为 n 次测量值，每次测量的误差为 $e_i = x_i - \theta$，按最小二乘法，误差累积为

$$L(\theta)=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}(x_i-\theta)^2,$$

求解 θ 使得 $L(\theta)$ 达到最小，正好是算术平均

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i.$$

由于算术平均是一个历经考验的方法，而以上的推理说明，算术平均是最小二乘法的一个特例，所以从另一个角度说明了最小二乘法的优良性，使我们对最小二乘法更加有信心。

最小二乘法发表之后很快得到了大家的认可接受，并迅速地在数据分析实践中被广泛使用。不过历史上又有人把最小二乘法的发明归功于高斯，这又是怎么一回事呢。高斯在 1809 年也发表了最小二乘法，并且声称自己已经使用这个方法多年。高斯发明了小行星定位的数学方法，并在数据分析中使用最小二乘法进行计算，准确地预测了谷神星的位置。

扯了半天最小二乘法，没看出和正态分布有任何关系啊，离题了吧？单就最小二乘法本身，虽然很实用，不过看上去更多的算是一个代数方法，虽然可以推导出最优解，对于解的误差有多大，无法给出有效的分析，而这个就是正态分布粉墨登场发挥作用的地方。勒让德提出的最小二乘法，确实是一把在数据分析领域披荆斩棘的好刀，但是刀刃还是不够锋利；而这把刀的打造后来至少一半功劳被归到高斯，是因为高斯不但独自地给出了造刀的方法，而且把最小二乘法这把刀的刀刃磨得无比锋利，把最小二乘法打造成了一把瑞士军刀。

高斯拓展了最小二乘法，把正态分布和最小二乘法联系在一起，并使得正态分布在统计误差分析中确立了自己的地位，否则正态分布就不会被称为高斯分布了。那高斯这位神人是如何把正态分布引入到误差分析之中，打造最小二乘法这把瑞士军刀的呢？

四、众里寻她千百度：误差分布曲线的确立

俄罗斯游行队伍里的正态分布标语

第三个故事有点长，主角是高斯和拉普拉斯，故事的主要内容是寻找随机误差分布的规律。

天文学是第一个被测量误差困扰的学科，从古代至十八世纪天文学一直是应用数学最发达的领域，到十八世纪，天文学的发展积累了大量的天文学数据需要分析计算，应该如何来处理数据中的观测误差成为一个很棘手的问题。我们在数据处理中经常使用平均的常识性法则，千百年来的数据使用经验说明算术平均能够消除误差，提高精度。算术平均有如此的魅力，道理何在，之前没有人做过理论上的证明。算术平均的合理性问题在天文学的数据分析工作中被提出来讨论：测量中的随机误差应该服从怎样的概率分布？算术平均的优良性和误差的分布有怎样的密切联系？

伽利略在他著名的《关于两个主要世界系统的对话》中，对误差的分布做过一些定性的描述，主要包括：