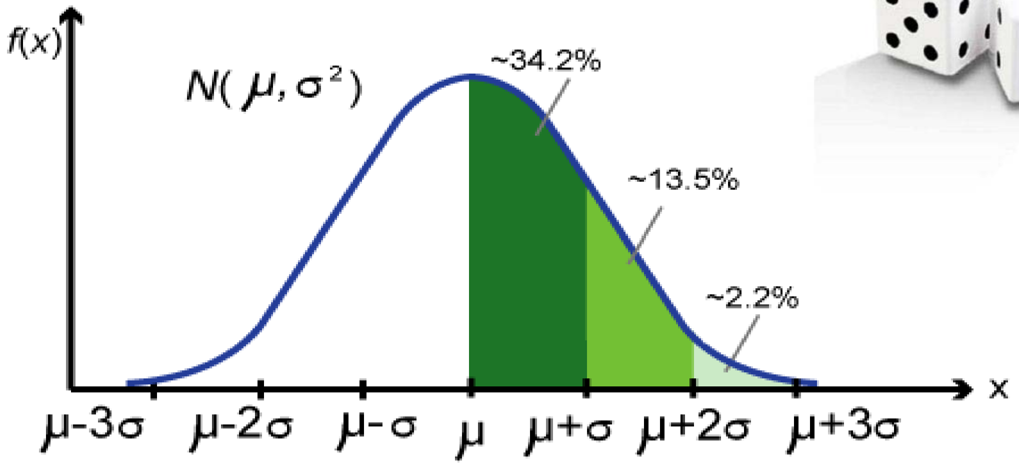
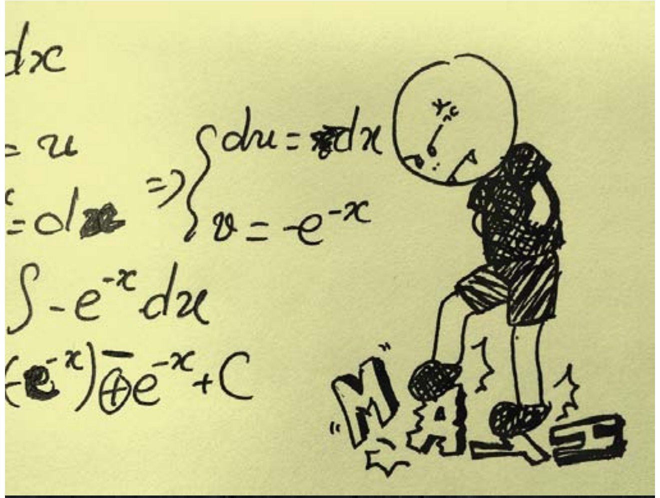
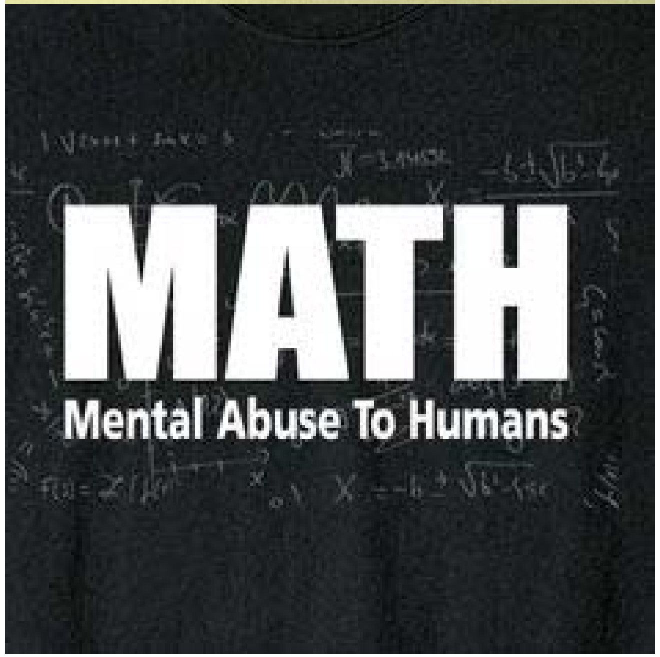
正态分布的前世今生

Rickjin(靳志辉)







小学生写诗吐槽数学

数学是死亡之源，
它像入地狱般痛苦。
它让孩子想破脑汁，
它让家长急得转圈。
它让校园死气沉沉，
它使生命慢慢离去。
生命从数学中走去，
一代代死得超快。
那是生命的敌人，
生命从数学中走去，
珍惜宝贵的生命吧。

数学 --- 冰冷的美丽 vs. 火热的思考

没有一种数学思想，以它被发现时的那个样子发表出来。一个问题被解决以后，相应地发展成一种形式化的技巧，结果使得**火热的思考变成了冰冷的美丽**。

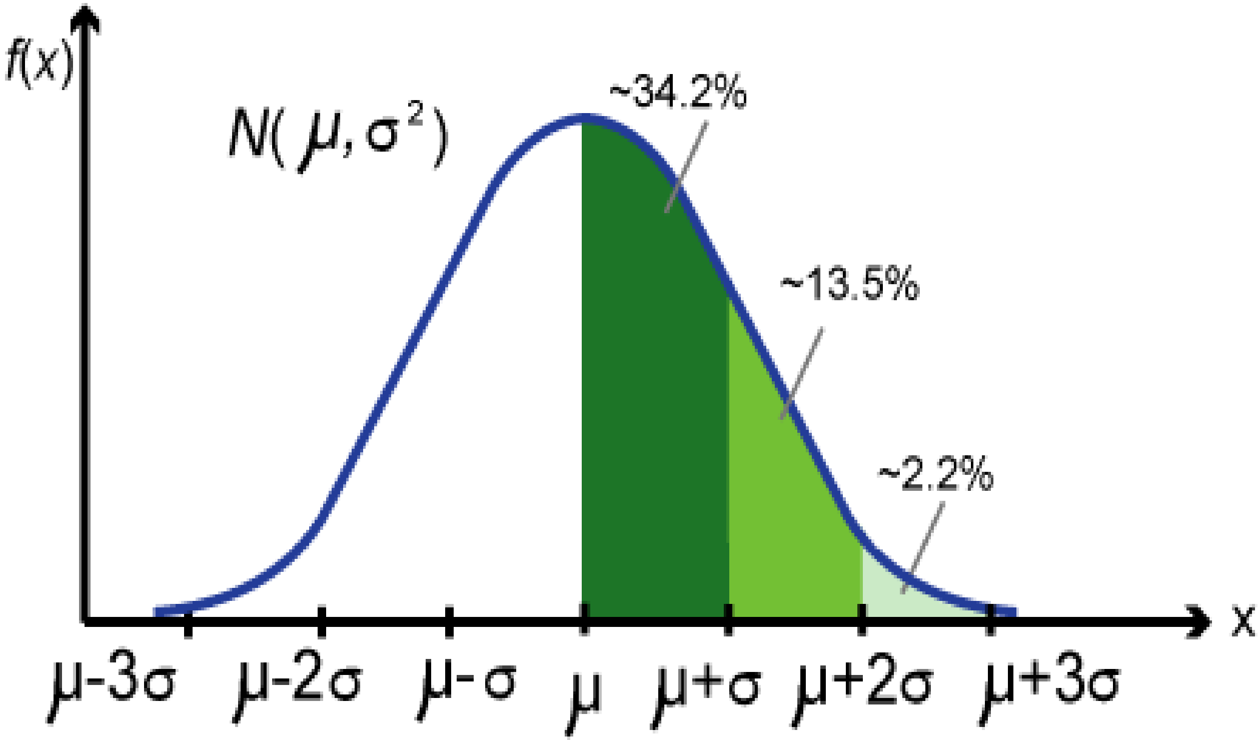
--- 国际数学教育委员会前主席 H. Freudenthal

在前辈数学家中欧拉对我的影响最大。主要原因在于，欧拉做了一些跟他才能相当的伟大数学家从没做过的事，即他解释了他是如何发现他的结论的。对此，我如获至宝。

---波利亚(George Pólya)

正态分布-- 熟悉的陌生人

这么怪异的曲线数学家们是如何找到的？



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

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

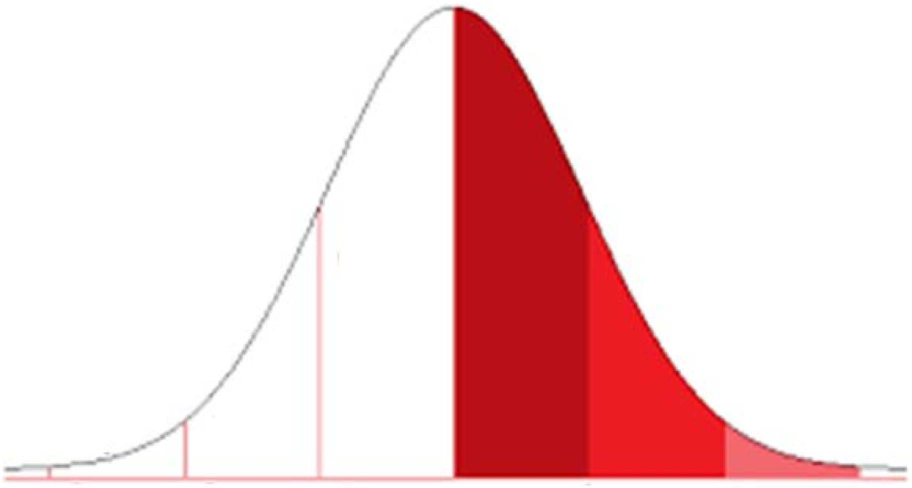
NORMAL

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

正态分布的前世今生

- 邂逅
- 瑞士军刀
- 众里寻她千百度
- 曲径通幽处，禅房花木深
- 华山论剑
- 开疆拓土，垂帘听政
- 大道至简，大美天成



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

Story 1：邂逅—正态曲线的首次发现

- 棣莫弗 (Abraham de Moivre, 1667-1754)
- 拉普拉斯 (Pierre-Simon Laplace 1749-1827)

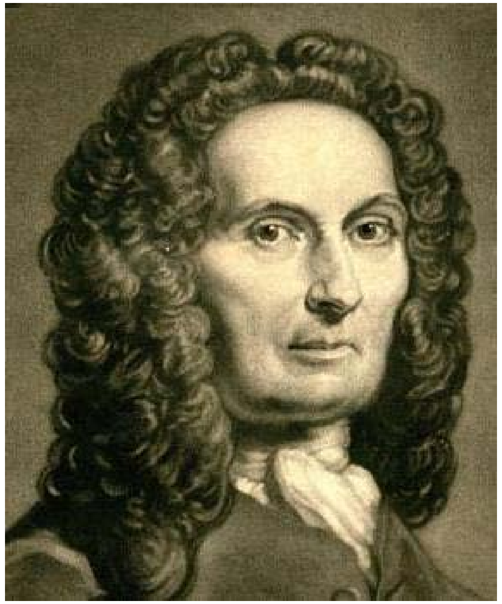
概率理论发源于赌博

- 惠更斯、帕斯卡、费马都是古典概率的奠基人。
- 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。

赌徒梅累向棣莫弗提出的问题

- 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，若 A 赢的局数记为 X，
- 若 X> np, 则 A 付给赌场 X-np 元，
- 否则B 付给赌场 np-X 元。

Q: 问赌场挣钱的期望值是多少?



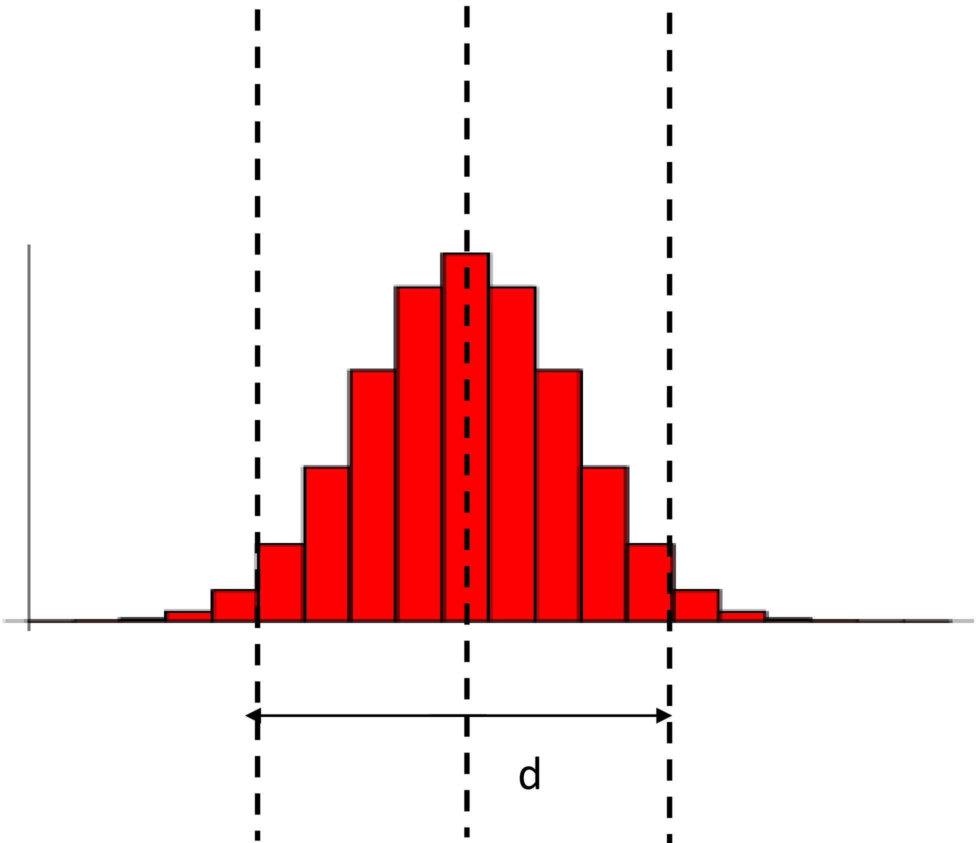
棣莫弗 (1667 ~ 1754)

棣莫弗开始研究二项分布的近似计算问题

$$X \sim B(n,p)$$

$$\mathrm{P}_d = P(|X - np| \leq d)$$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$



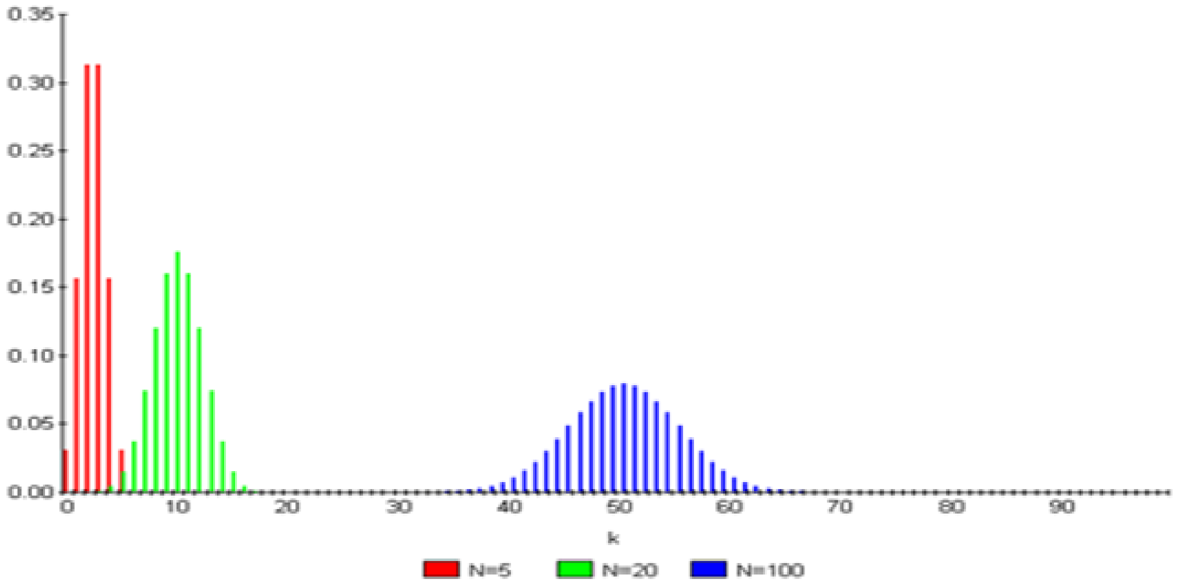
借助斯特灵公式，邂逅正态分布曲线

$$\mathrm{P}_d = P(|X - np| \leq d)$$

处理特例 p=1/2

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

$$P\left(\left|\frac{X}{n} - \frac{1}{2}\right| \leq \frac{c}{\sqrt{n}}\right) \sim \int_{-2c}^{2c} \boxed{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}} dx$$



概率论大牛拉普拉斯
研究独立随机变量求和的概率的近似计算

- 二项分布 $p \neq ½$ 的研究

- Laplace 研究彗星轨道倾角遇到的问题

$$S_n = X_1 + X_2 + ... + X_n$$

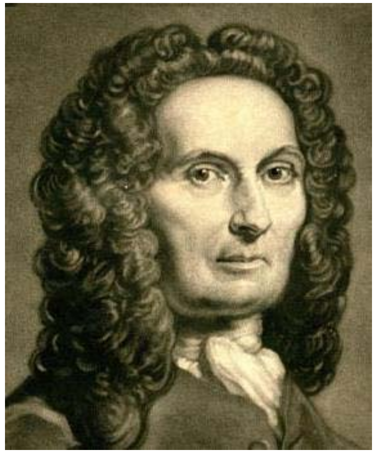
$$P(a \leq S_n \leq b) = ?$$

拉普拉斯 (1749~1827)

正态分布曲线的首次发现 --- 中心极限定理 (Central Limit Theorem, 1775)

定理 0.2.1 (棣莫弗 -拉普拉斯中心极限定理) 设随机变量 $X_n(n = 1, 2, \cdots)$ 服从参数为 n, p 的二项分布，则对任意的 x, 恒有

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



棣莫弗 (1667 ~ 1754)

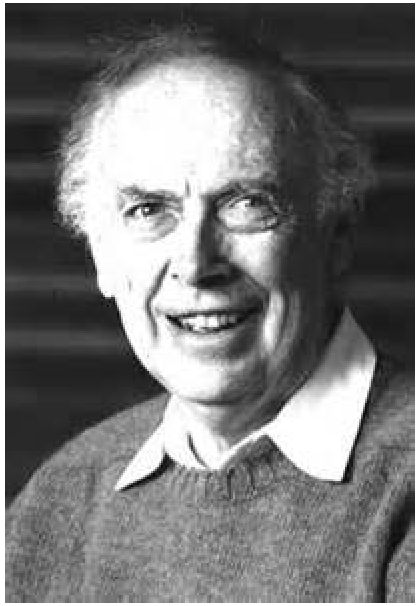
拉普拉斯 (1749~1827)

科学发现的逻辑

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一样，按照直接了当合乎逻辑的方式进行的。

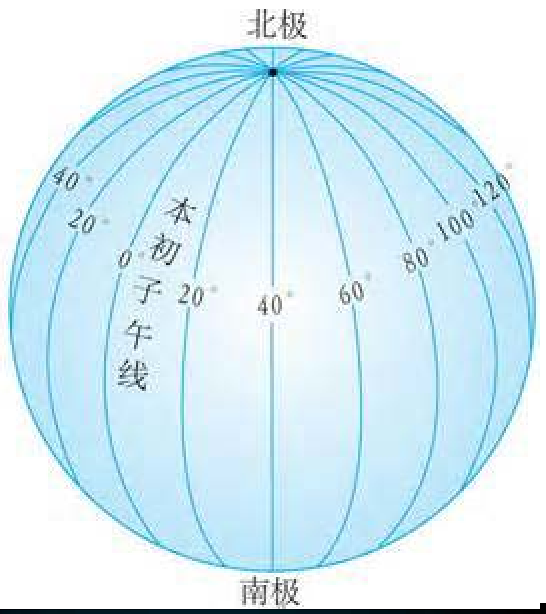
-- DNA 双螺旋的发现者 James D. Watson



Story 2: 最小二乘法，打造数据分析的瑞士军刀

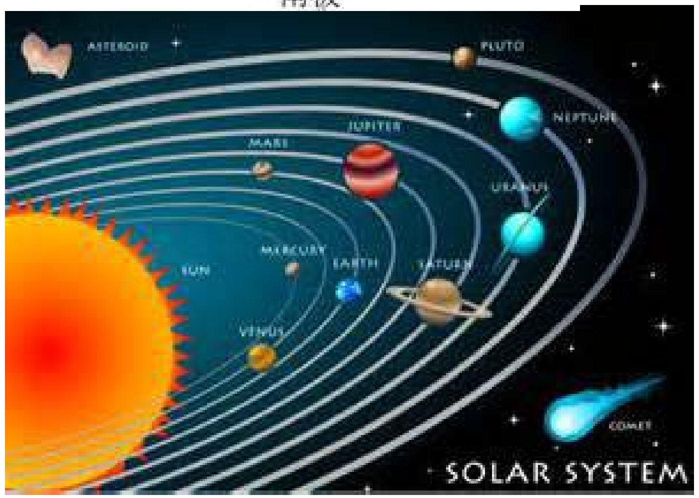
- 欧拉(Leonhard Euler, 1707-1783)
- 拉普拉斯 (Pierre-Simon Laplace 1749-1827)
- 勒让德(Adrien-Marie Legendre, 1752–1833)
- 高斯(Carl Friedrich Gauss, 1777-1855)

天文学是数学最发达的领域，19世纪数据分析成为重要的工作



- 勒让德 --- 子午线度量
- 拉普拉斯 --- 慧星轨道倾角计算
- 高斯 --- 行星轨道计算

- 大量的天文学和测地学数据需要处理
- 同一个指标有多项测量结果，存在测量误差



Q: 如何有效的处理测量误差 ？

天文和测地学中的问题建模

- x,y 为观察值， beta 为需要估计的参数

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\
&\cdots \\
y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np}
\end{aligned}
$$

挑战： n >> p
线性矛盾方程组， 无法用直接解方程的方式处理

大数学家拉普拉斯& 欧拉的努力

- 线性矛盾方程
- 把多个方程合并，简化为 p 个不矛盾方程进行求解
- 方法缺乏通用性，都以失败告终

拉普拉斯 (1749~1827)

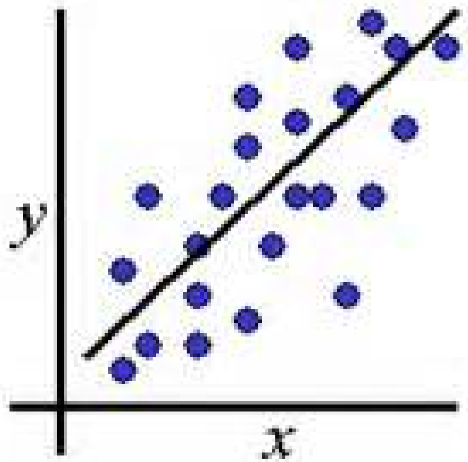
欧拉(1707-1783)

勒让德发表最小二乘法（1805）

累积误差 = $\sum$(观测值 - 理论值)^2

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

勒让德(1752–1833)



最小二乘法成为数据分析的瑞士军刀

- 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

假设真值为 $\theta, x_1, \cdots, x_n$ 为 n 次测量值

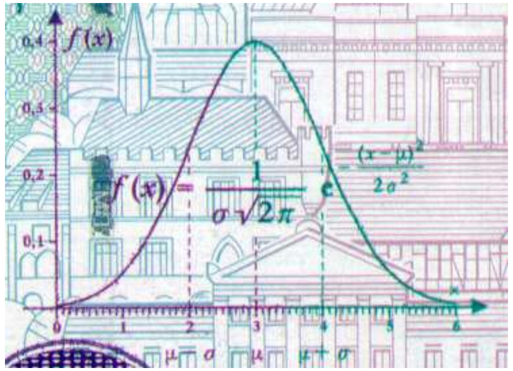
$$L(\theta) = \sum^{n} e_i^2 = \sum^{n} (x_i - \theta)^2$$

求解 θ 使得 $L(\theta)$ 达到最小，正好是算术平均 $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

高斯 PK 勒让德 --- 最小二乘法发明权之争

- 高斯1809 年发表基于正态分布的最小二乘法
- 宣称自 17xx 年已经开始使用该方法
- 高斯的方法优于勒让德，给最小二乘法添加了锋利的刀刃
- 高斯威望巨大，研究严谨
  “few, but ripe”

正态分布被印在德国的货币上



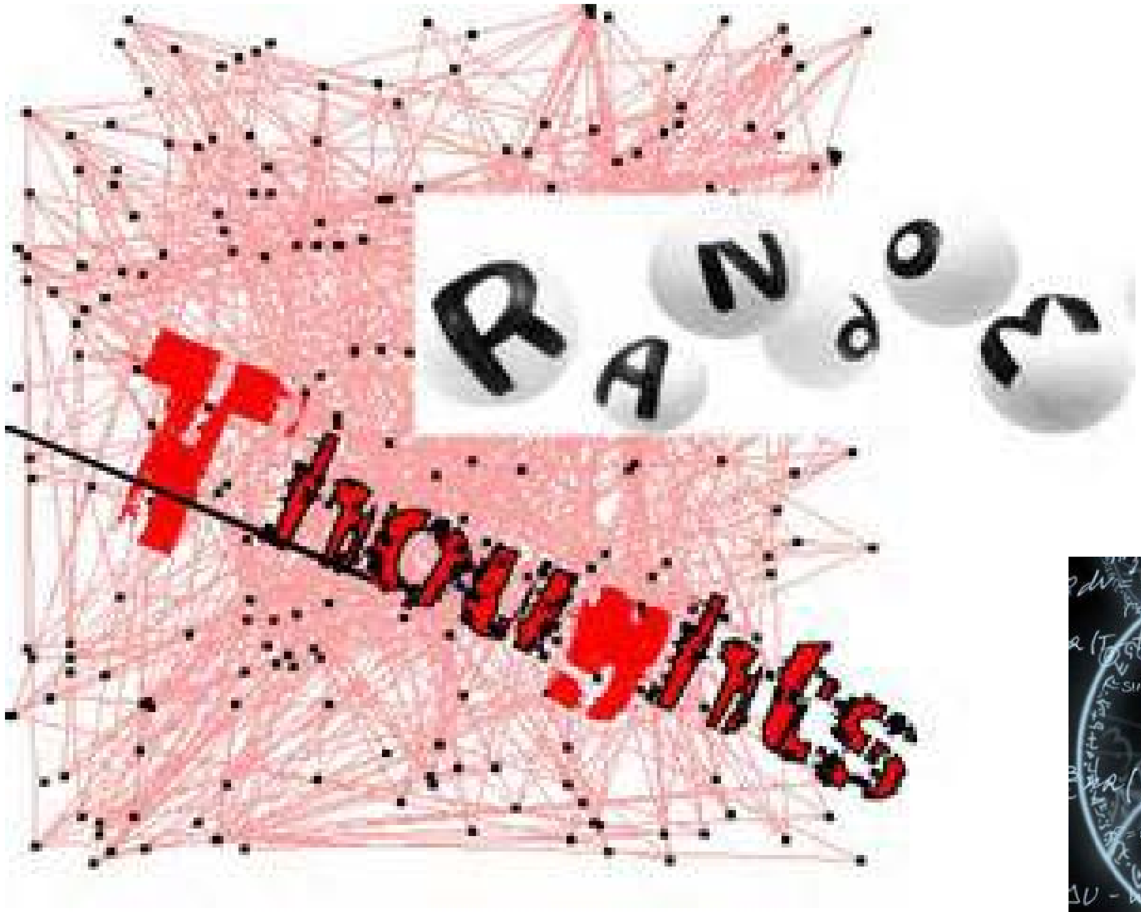
Story 3: 众里寻她千百度, 误差分布曲线的确立

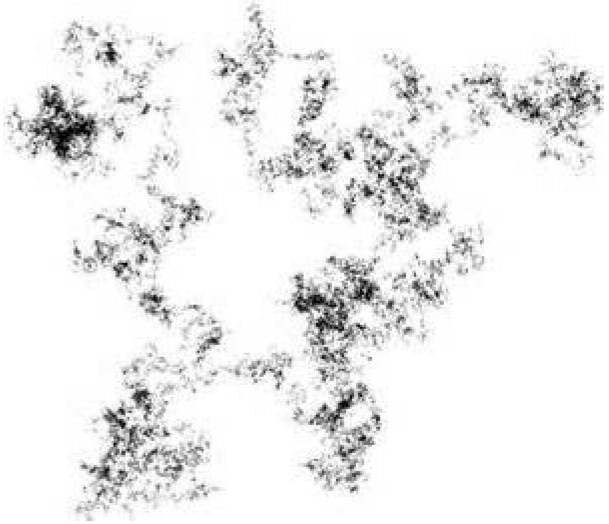
- 伽利略
- 辛普森
- 拉普拉斯
- 高斯

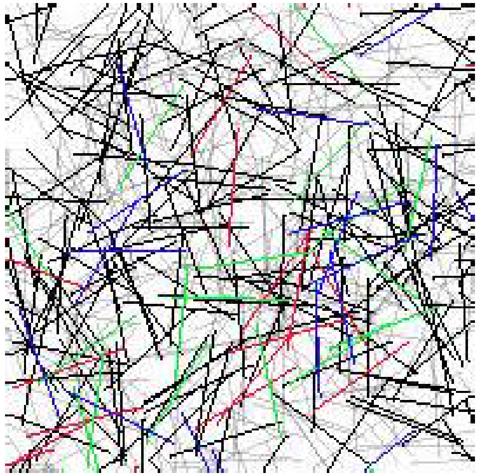
天文学的发展提出了对随机测量误差的分析需求

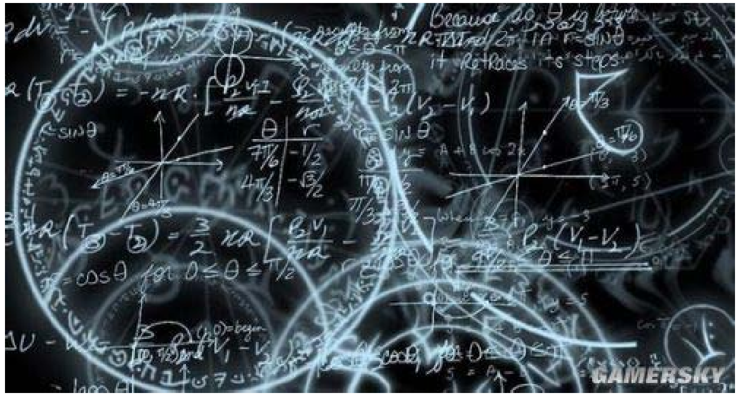
- 18世纪，天文学积累了大量的数据观测误差如何处理？
- 经验说明算术平均能够消除误差，提高精度，但是道理何在？

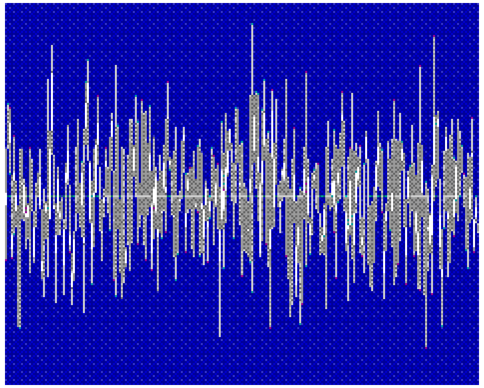
随机测量误差应该服从怎样的概率规律?











随机误差分析的先锋：定性描述与定量分析

- 伽利略
  - 误差是对称分布的
  - 大的误差出现频率低，小的误差出现频率高
  - 误差函数 f(x) 关于0对称分布，概率密度随 |x|增加而减小

- 辛普森
  - 第一次从概率论的角度论证算术平均的优良性。

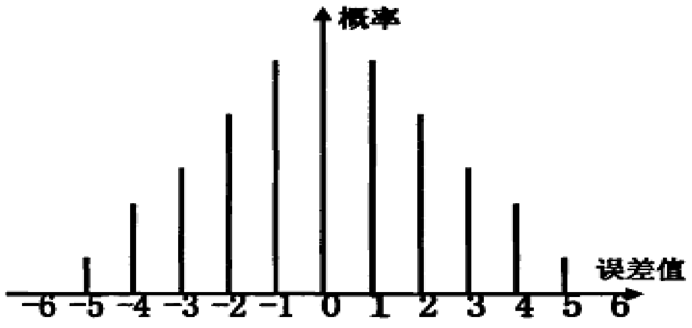


图 1 概率分布

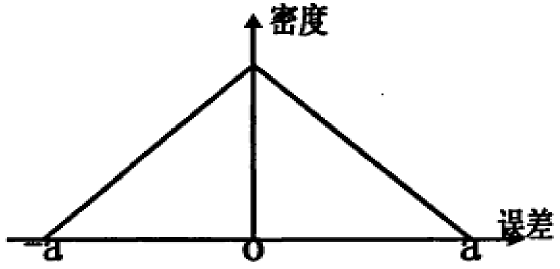


图 2 三角形分布

拉普拉斯提出的误差概率分布

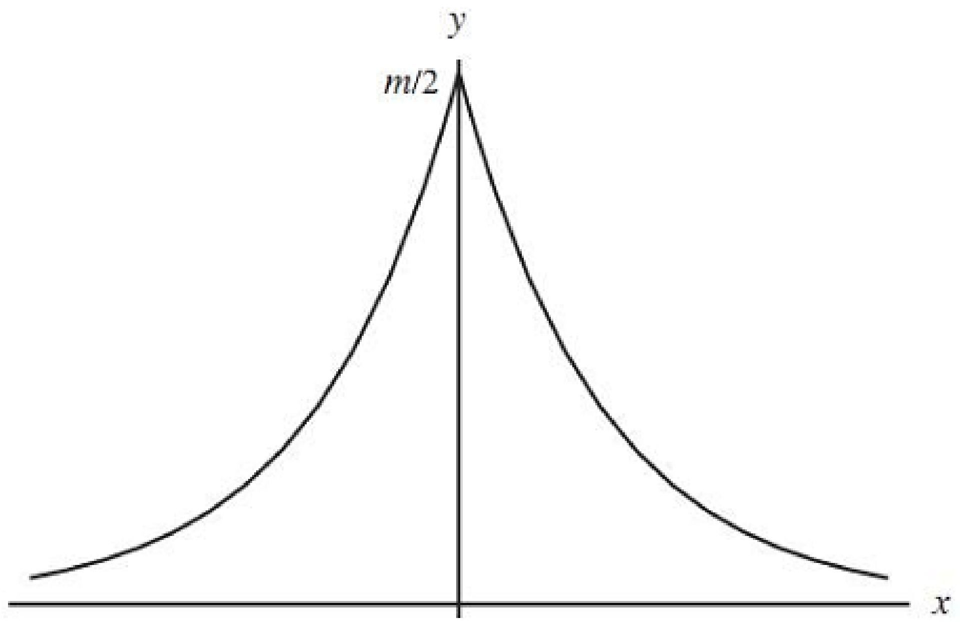


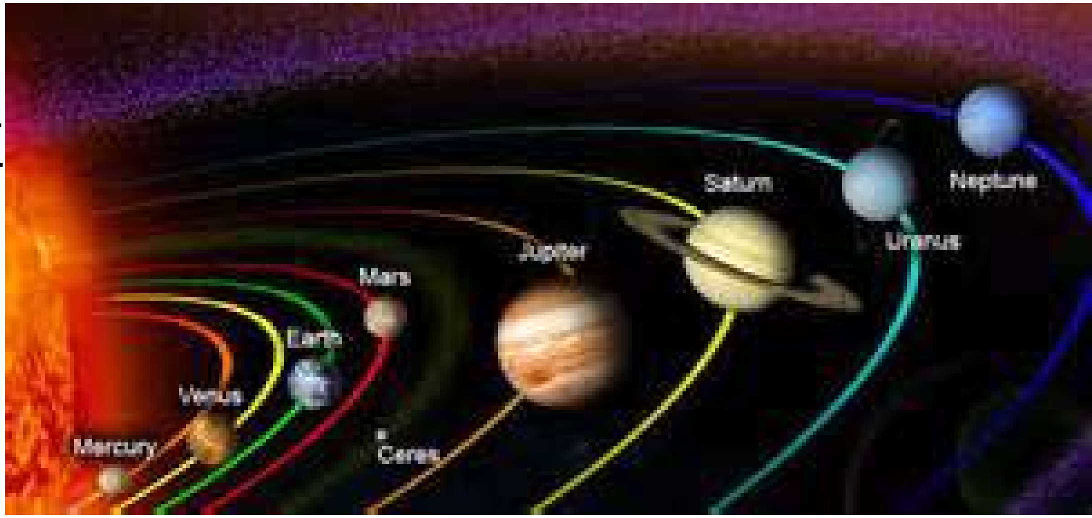
图 6: 拉普拉斯的误差分布曲线

$$f(x) = \frac{m}{2} e^{-m|x|}$$

高斯的介入--- 谷神星（Ceres)事件

- 1801年1月，一颗光度8等的新星出现
- 出现6个星期，扫过八度角后，阳光下无法观测
- 观测数据有限，难以计算出轨道
- 彗星还是行星？

- 高斯创立了新的行星轨道的计算方法，给出了谷神星的轨道
- 1801年12月，在高斯预言的的时间和位置重新被发现
- 数据分析方法为基于正态分布的最小二乘方法
- 1809 年发表



高斯推导随机误差的分布

- 基本假设
  - 设真值为θ, n 次独立测量值为 xi
  - 随机误差 ei = xi–θ 的概率分布为 f(x)
- 首次引入了极大似然估计的思想

$$L(\theta) = L(\theta; x_1, \cdots, x_n) = f(e_1)\cdots f(e_n)$$
$$= f(x_1 - \theta)\cdots f(x_n - \theta)$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

高斯逆向思维揣测上帝的意图

误差分布导出的极大似然估计 ＝ 算术平均

$$\hat{\theta} = \arg\max L(\theta) = \frac{X_1 + \cdots + X_n}{n}$$

Q : 什么样的随机误差函数f(x) 满足以上条件？

A : 数学上可以证明只有如下函数满足要求

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

通过正态分布&极大似然思想导出最小二乘法

$$e_1 = y_1 - (\beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p})$$

$$e_2 = y_2 - (\beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p})$$

$$\cdots$$

$$e_n = y_n - (\beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np})$$

$$e_i \sim N(0, \sigma^2)$$

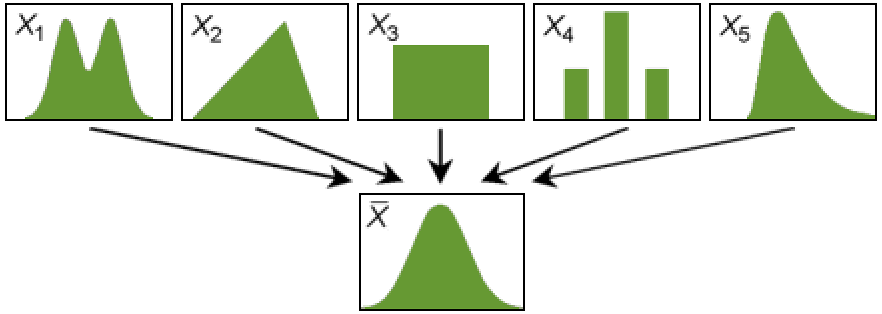
$$(e_1, e_2, \cdots e_n) \sim (\sqrt{2\pi}\sigma)^{-n} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i{}^2)$$

要使得这个概率最大，必须使得 $\sum_{i=1}^{n} e_i^2$ 取最小值

拉普拉斯元误差理论(1810)

- 高斯的推导，鸡生蛋还是蛋生鸡？
  - 算术平均 + 极大似然估计 → 误差为正态分布
  - 误差为正态 → 最小二乘 → 算术平均

- 误差是由大量的、由种种原因产生的元误差叠加而形成
- 由中心极限定理，误差必然服从正态分布



正态分布历史地位的确定

- 棣莫弗导出正态曲线 (1733)
- 棣莫弗-拉普拉斯中心极限定理 (1785)
- 拉普拉斯给出一般性的中心定理描述 (18xx)
- 高斯推导出随机误差为正态分布(1809)
- 拉普拉斯用中心极限定理给出元误差解释(1810)

正态分布为何不称为棣莫弗分布？而是被称为高斯分布？

- Stigler’s Law

  绝大多数成果的冠名，大都不是历史上首位发现者的名字

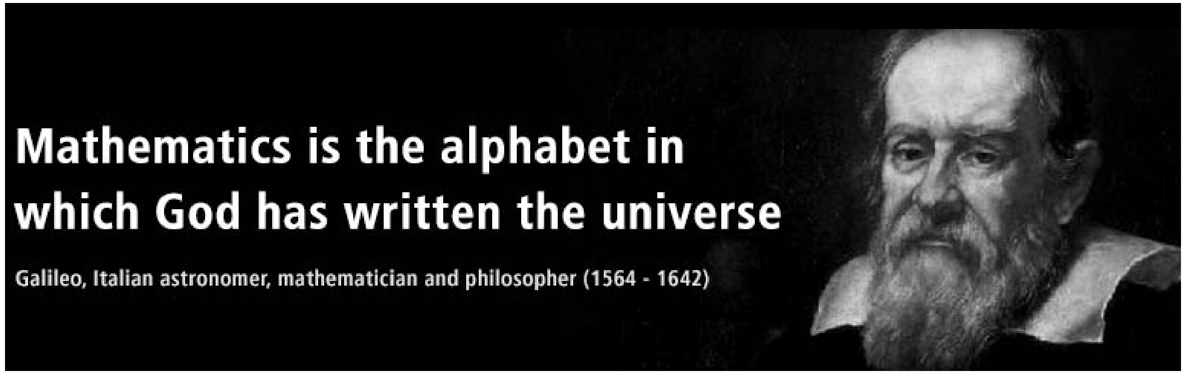
- 是高斯赋予了正态分布统计学上的意义

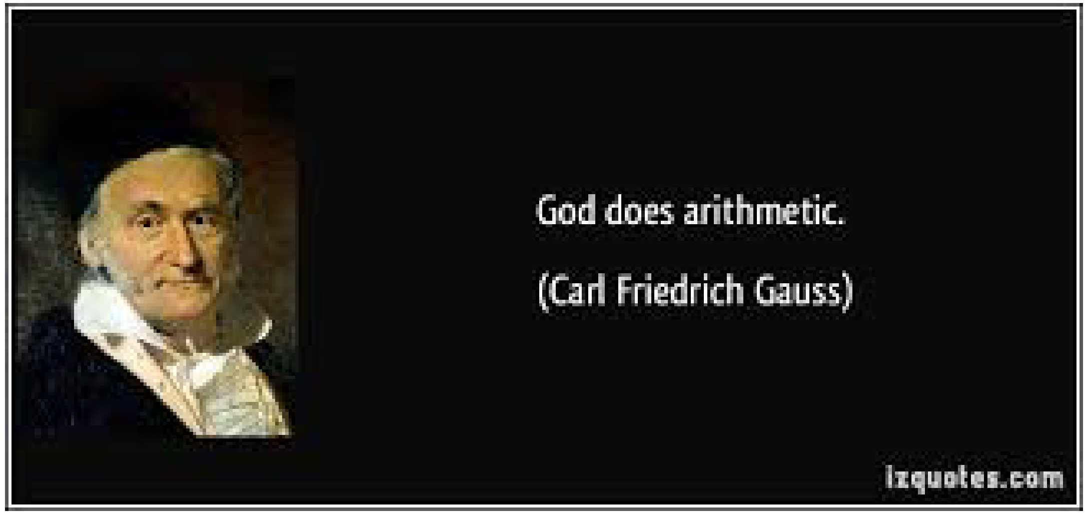
正态分布冠名权之争

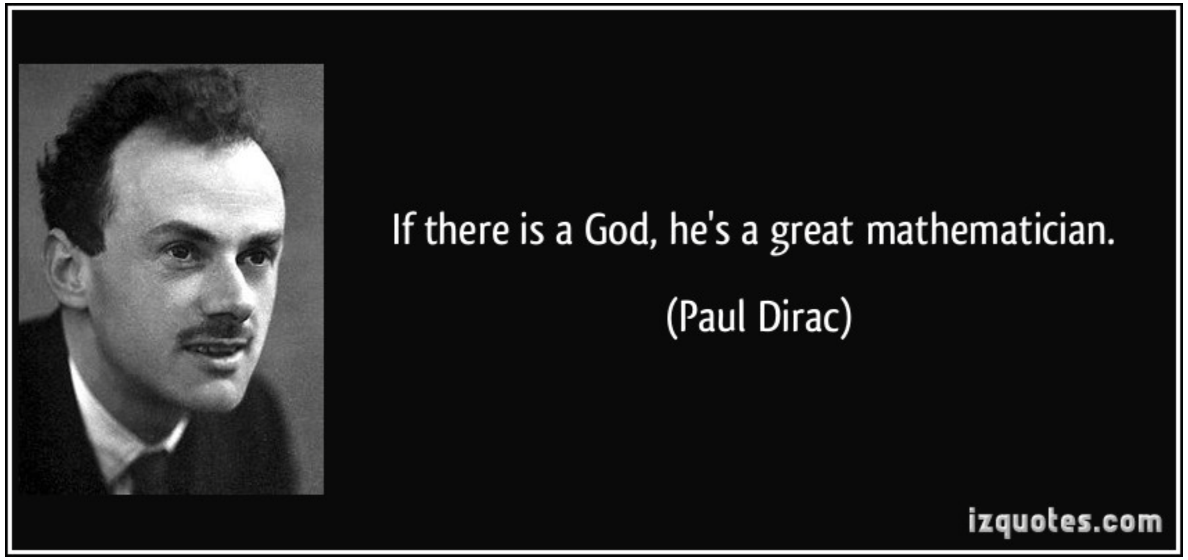
- 正态分布在不同国家的名称
  - 法国，拉普拉斯分布
  - 德国，高斯分布
  - 其他国家，拉普拉斯-高斯分布

- 正态分布的名称由来
  - 后来法国的大数学家庞加莱建议改用正态分布这一中立名称
  - 统计学家卡尔·皮尔森(Karl Pearson) 使得这个名称被广泛接受
  - 由于高斯的贡献以及在数学界的威望，许多人使用了高斯分布

God Is A Mathematician







要成为一个好的数学家,……,
你必须首先是一个好的猜想家

--- 波利亚(George Pólya)

Story 4: 曲径通幽处，禅房花木深

条条曲径通正态

- 棣莫弗 (1733)
- 高斯 (1809)
- 拉普拉斯(1810)

- 赫歇尔(1850)和麦克斯韦(1860)
- 兰登(1941)
- 香农(1948)

高斯(1809)的推导

- 准则

  极大似然估计 = 算术平均

$$\hat{\theta} = \arg\max L(\theta) = \frac{X_1 + \cdots + X_n}{n}$$

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

只有正态分布 N(0, σ^2) 满足要求

赫歇尔(1850)和麦克斯韦(1860)的推导

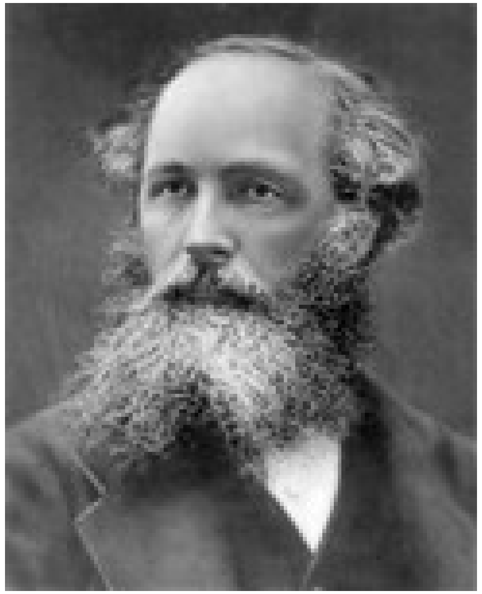
- 准则
  - 随机误差在正交的方向上相互独立
  - 误差的概率分布在空间上具有旋转对称性，即误差的分布和角度没有关系

$$p(x,y)=\frac{\alpha}{\pi}e^{-\alpha(x^2+y^2)}$$

- 三维空间分子运动速度的分布
  - 麦克斯韦-波尔兹曼气体速率分布定律

$$\rho(v_x,v_y,v_z)\propto \exp\{-\alpha(v_x^2+v_y^2+v_z^2)\}$$

$$\begin{aligned}F(v) &= \left(\frac{m}{2\pi kT}\right)^{3/2}e^{-\frac{mv^2}{2kT}}\\ &= \left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv_x^2}{2kT}}\times\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv_y^2}{2kT}}\times\left(\frac{m}{2\pi kT}\right)^{1/2}e^{-\frac{mv_z^2}{2kT}}\end{aligned}$$



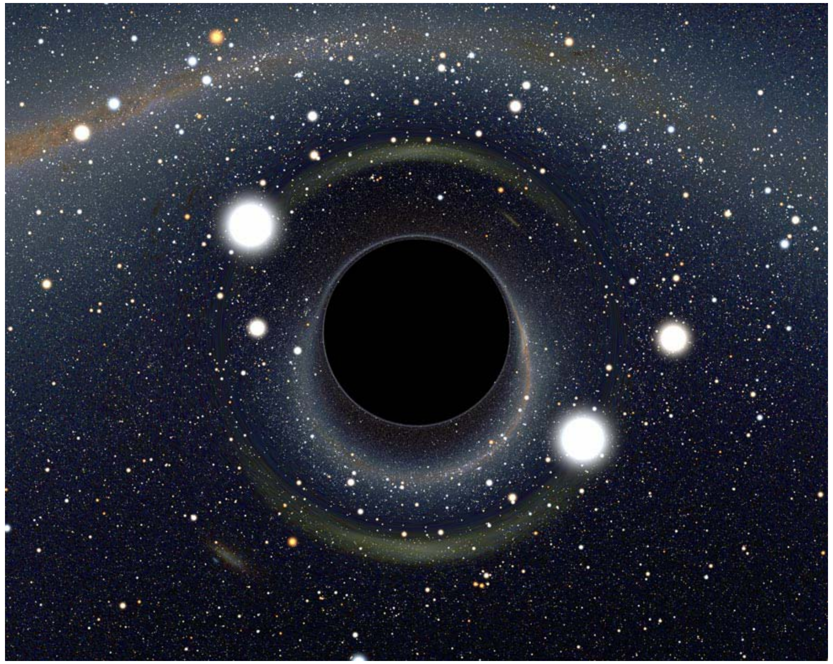
兰登(1941)的推导

- 准则：
  - 随机噪声具有稳定的分布模式
  - 累加一个微小的随机噪声，不改变其稳定的分布模式，只改变分布的层级(用方差度量)

$$x \sim p(x;\sigma^2) \qquad e \sim q(e)$$

$$x' = x + e$$

$$x' \sim p(x;\sigma^2 + \mathrm{var}(e))$$

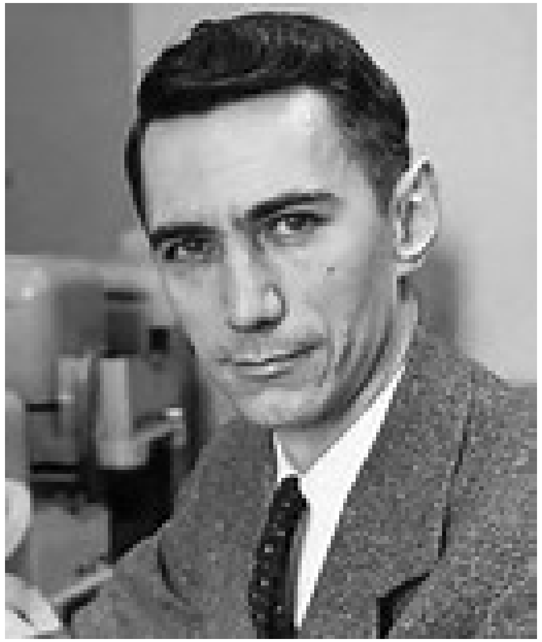


正态黑洞

香农的推导(1948)

- 准则
  - 给定均值和方差，求熵最大的概率分布

$$H(p) = -\int p(e) \log p(e) de$$



- 给定均值和方差，正态分布具有最大熵
- 从信息论的角度证明了正态分布的优良性

禅房花木深

Physicists believe that the Gaussian law has been proved in mathematics while mathematicians think that it was experimentally established in physics.

物理学家认为高斯分布已经在数学上得到证明，而数学家则认为高斯分布在物理试验中得到确认。

— 庞加莱(Henri Poincaré)

Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact.

— Henri Poincaré

Story5: 论剑中心极限定理

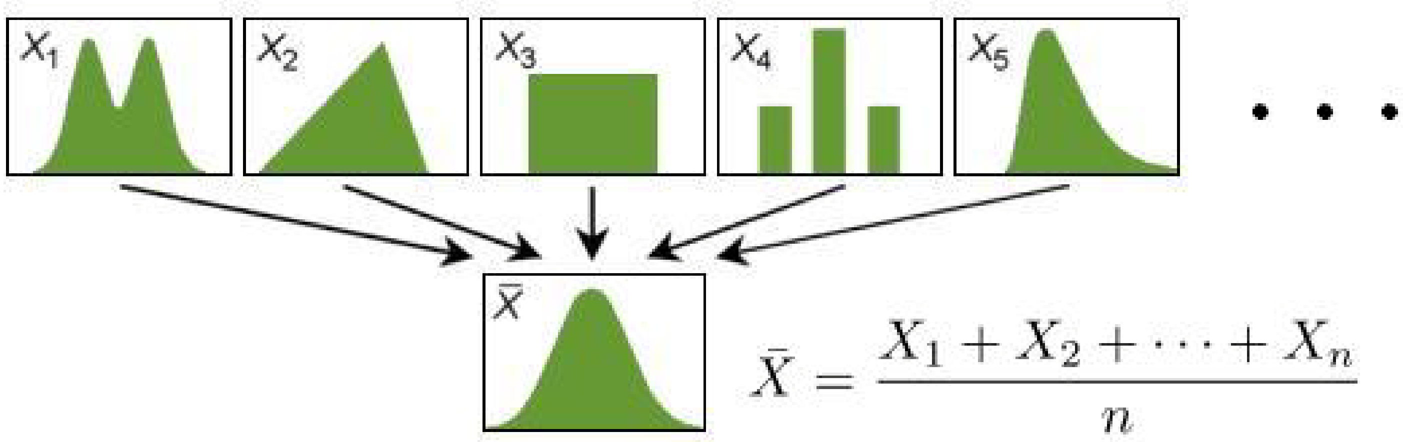


图 8: 华山论剑

拉普拉斯的逼近

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，$a>0$，则有

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

中心极限定理的严格证明

- 泊松(Siméon Denis Poisson, 1781-1840)
- 狄利克莱(Gustav Dirichlet, 1805-1859)
- 柯西(Augustin-Louis Cauchy,1789-1857)
- 贝塞尔(Friedrich Bessel, 1784-1846)
- 切比雪夫(Chebyshev, 1821-1894)
- 马尔可夫(Markov, 1856-1922)
- 李雅普诺夫(Lyapunov, 1857-1918)

中心极限定理的证明犹如蜘蛛补洞

李雅普诺夫1901 年给出了第一个在一般条件下的严格证明

中心极限定理的一般形式

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

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

Q: 中心极限定理成立的充要条件是什么？

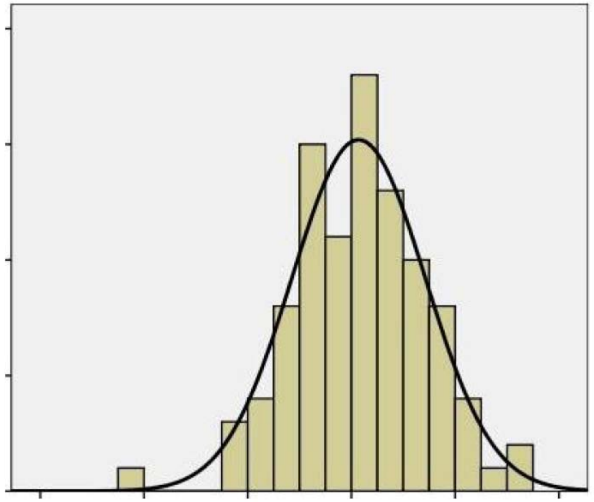
Story 6: 开疆拓土，垂帘听政

- 正态分布被引入自然科学和社会科学领域
- 统计学三大分布以及小样本理论

开疆拓土---正态分布进入社会科学领域

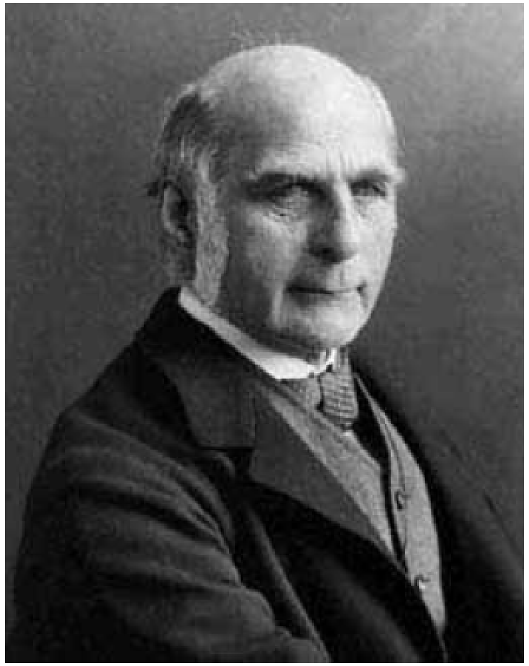
- 测量数据和统计数据有区别吗？
- 凯特勒 是19世纪最有影响力的统计学家之一，他倡导并身体力行将正态分布用于连续数据分析

凯特勒（Adolphe Quetelet, 1796-1874）

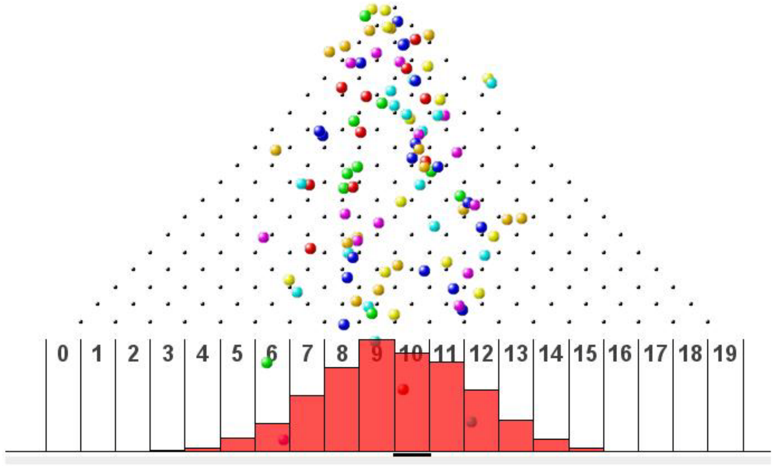


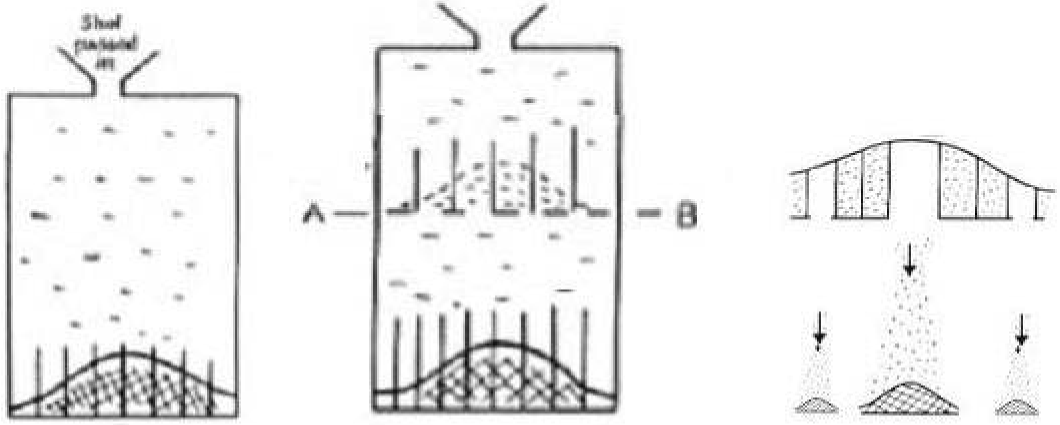
把一批数据是否能很好地拟合正态分布，作为判断该批数据是否同质的标准

开疆拓土---生物统计的兴起

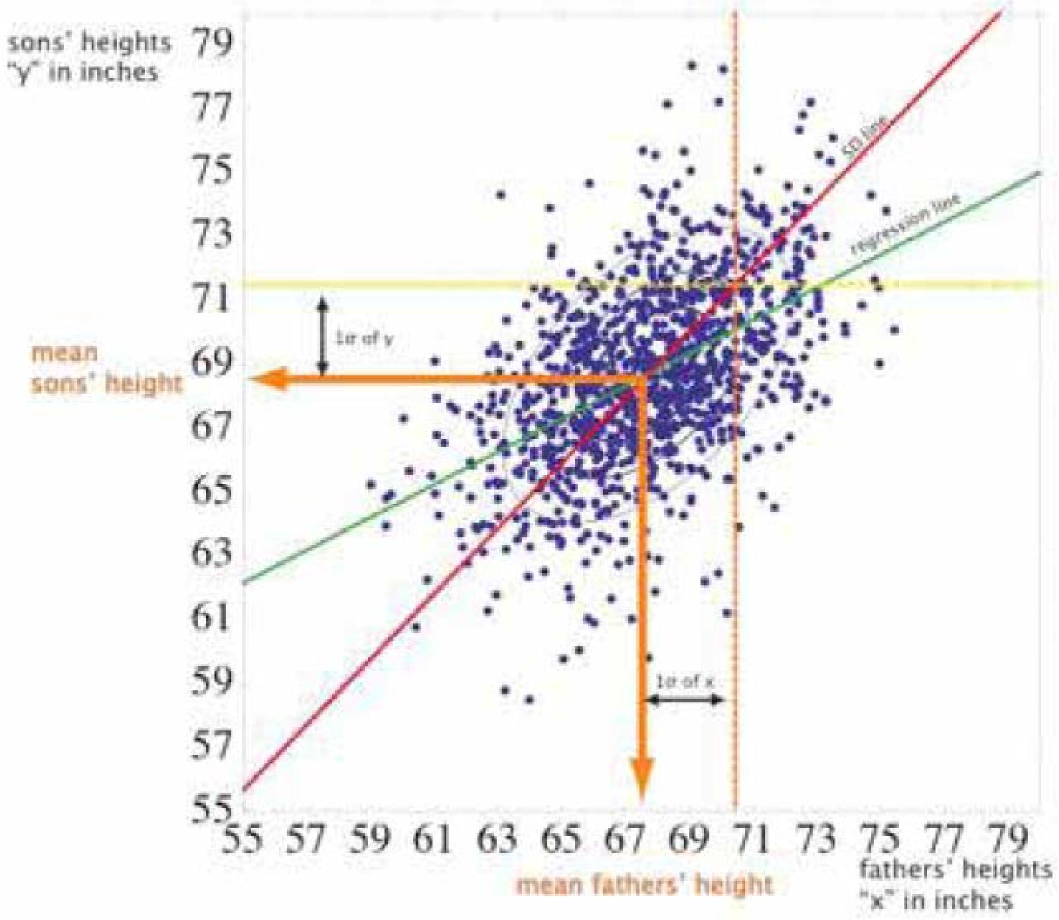


高尔顿（Francis Galton, 1822-1911）



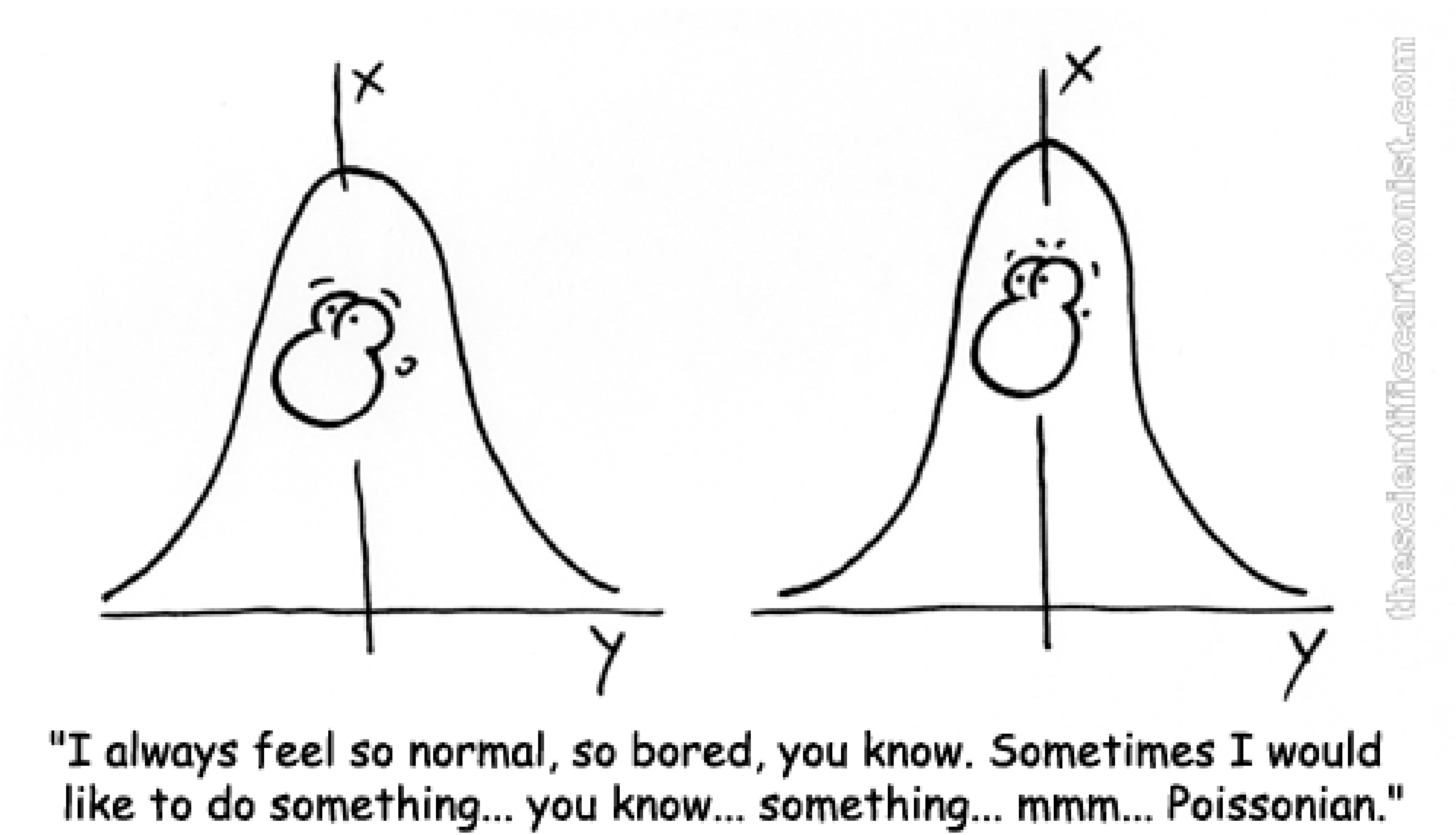


正态分布在回归分析中的应用

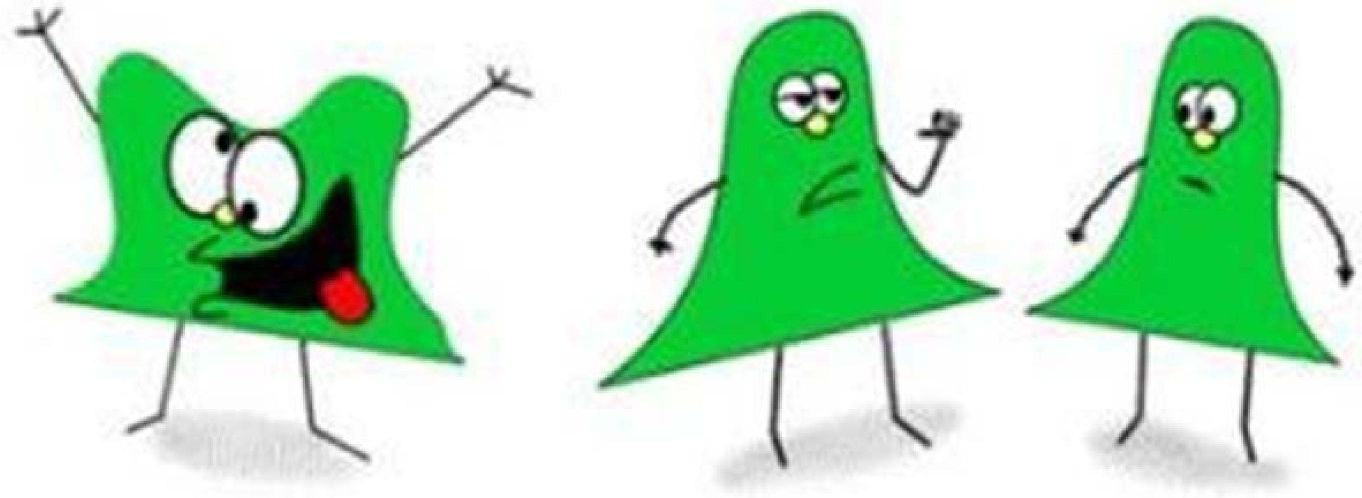


儿子与父亲的身高回归线

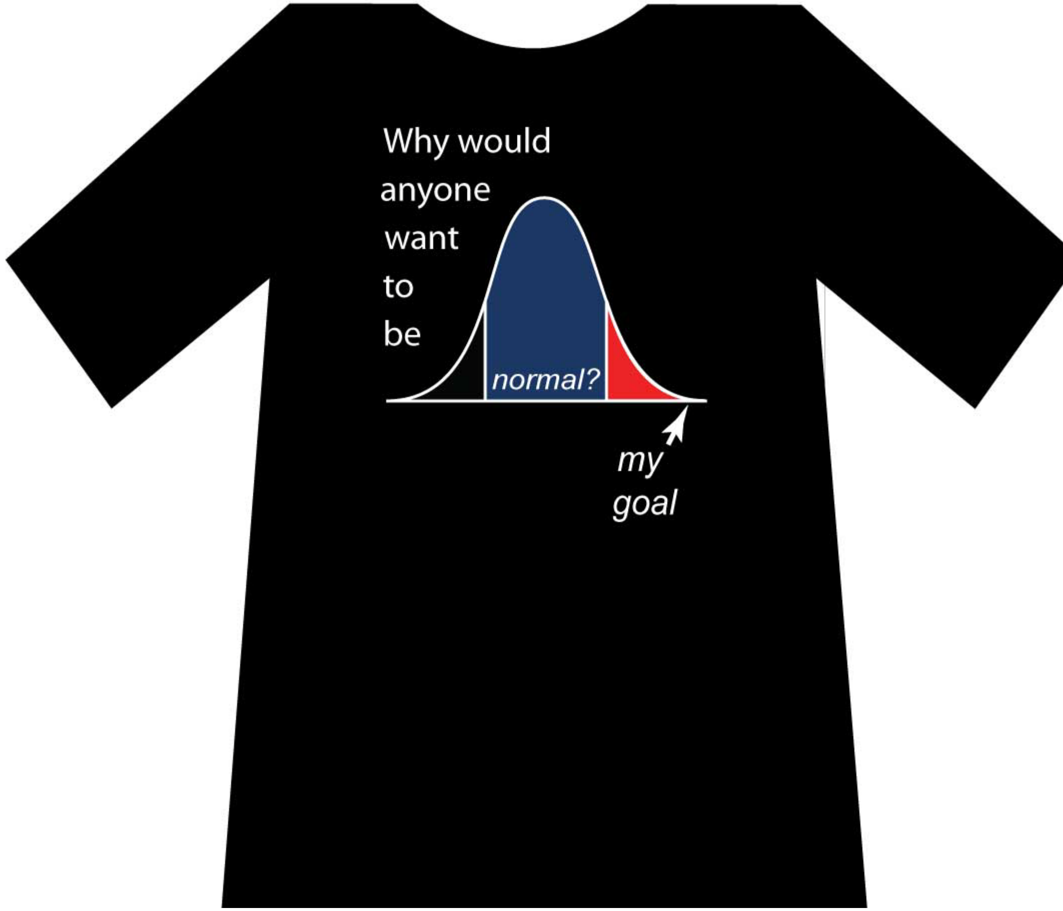
正态分布广泛应用到滥用



Normal vs. Abnormal



“ KEEP YOUR EYE ON THAT GUY, TOM,
HE’S NOT, YOU KNOW... NORMAL ! ”



现代数理统计三剑客

Karl Pearson

χ^2

W. S. Gosset

t

R. A. Fisher

F

垂帘听政 --- 小样本统计学的蓬勃发展

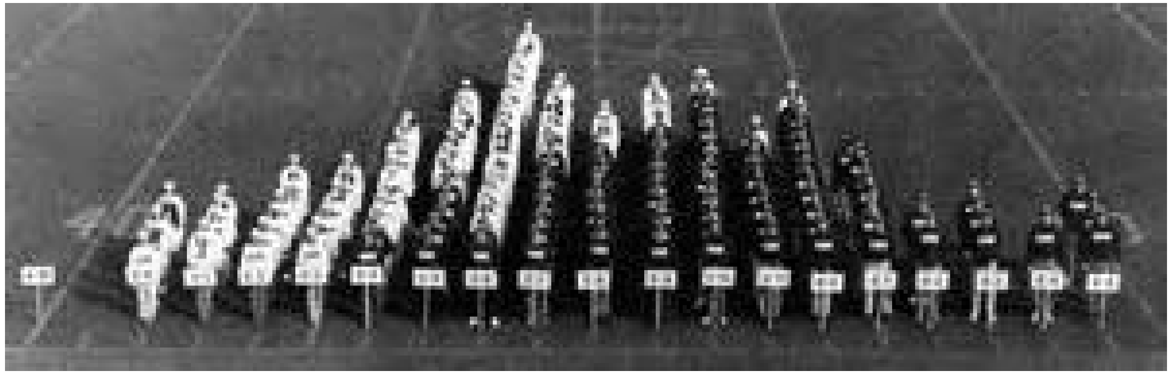
$$X_i \sim N(0,1), Y_j \sim N(0,1)$$

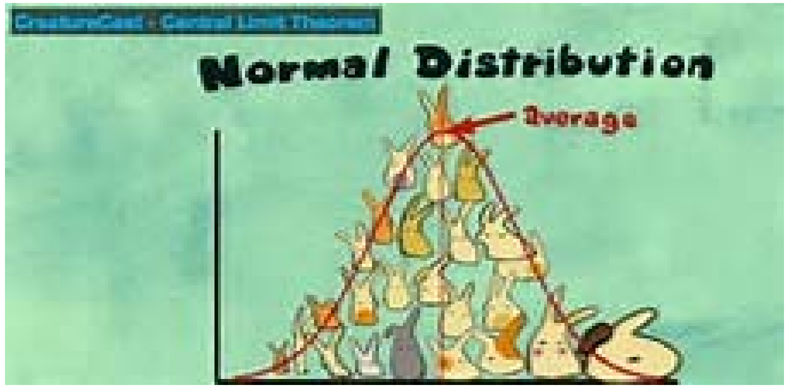
1. $\chi_n^2 = X_1^2 + \cdots + X_n^2$

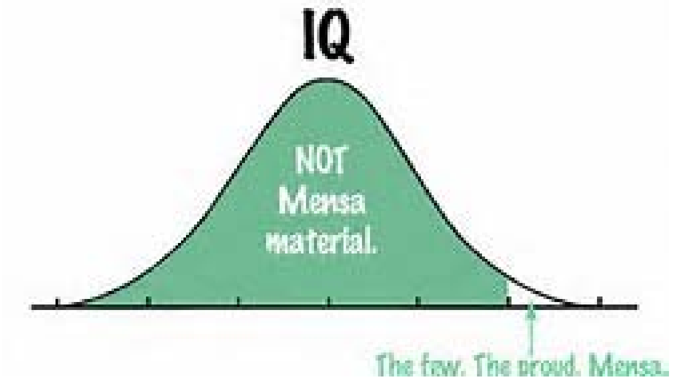
2. $t = \dfrac{Y_1}{\sqrt{\frac{X_1^2+\cdots+X_n^2}{n}}}$

3. $F = \dfrac{\frac{X_1^2+\cdots+X_n^2}{n}}{\frac{Y_1^2+\cdots+Y_m^2}{m}}$

Story 7：大道至简，大美天成







Q: 为何正态分布无处不在 ？

Q: 为什么实践中正态分布经常成功 ？

因为正态分布具有非常好的数学性质

- **正态分布稳定**，在各种操作之下形式不变
  - 两个正态分布密度的乘积还是正态分布
  - 两个正态分布密度的卷积还是正态分布
  - 正态分布的傅立叶变换还是正态分布
  - 中心极限定理保证了多个随机变量的求和效应将导致正态分布
- **正态分布具有最大熵**
  - 很多时候我们其实没有任何的知识知道数据的真实分布是什么，但是一个分布的均值和方差往往是相对稳定的，
  - 因此按照最大熵的原理，我们应该选择在给定的知识的限制下，选择熵最大的概率分布，而这就恰好是正态分布。

正态魅影

The
normal
law of error
stands out in the
experience of mankind
as one of the broadest
generalisations of natural
philosophy. It serves as the
guiding instrument in researches
in the physical and social sciences
and in medicine, agriculture and engineering.
It is an indispensable tool for the analysis and the
interpretation of the basic data obtained by observation and experiment.

W J Youden

正态魅影

我几乎不曾见过像误差呈正态分布这么美妙而激发人们无穷想象的宇宙秩序。如果古希腊人知道这条曲线，想必会给予人格化乃至神格化。它以一种宁静无形的方式在最野性的混乱中实施严厉的统治。暴民越多，无政府状态越显现，它就统治得越完美。他是无理性世界中的最高法律。当我们从混沌中抽取大量的样本，并按大小加以排列整理时，那么总是有一个始料不及的美妙规律潜伏在其中。

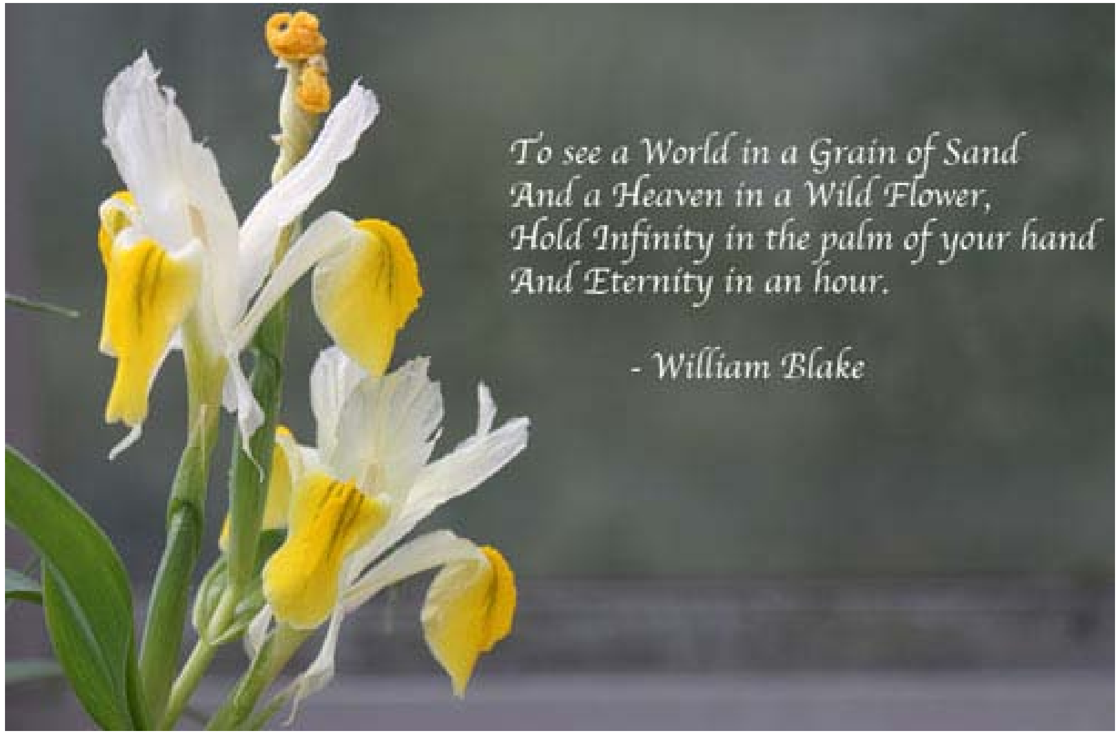
--- 高尔顿

我接触到正态分布之后马上被他深深的吸引，我感到难以相信，这个来自经验直方图和赌博游戏的规律，居然会成为我们日常生活数学的一部分

--- 概率学家卡克

大道至简，大美天成

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$



统计学之用

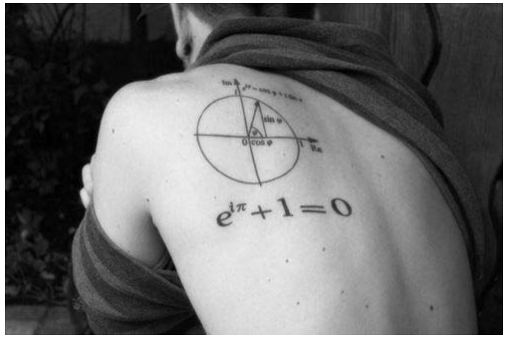
All knowledge is, in the final analysis, history.
All sciences are, in the abstract, mathematics.
All methods of acquiring knowledge are, essentially, through statistics.

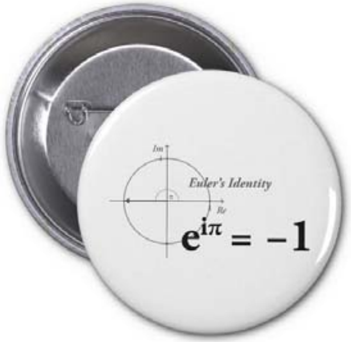
在终极的分析中，一切知识都是历史；
在抽象的意义下，一切科学都是数学；
在理性的基础上，所有的判断都是统计学。

— C. R. Rao

美丽的数学

$$e^{i\pi}+1=0$$

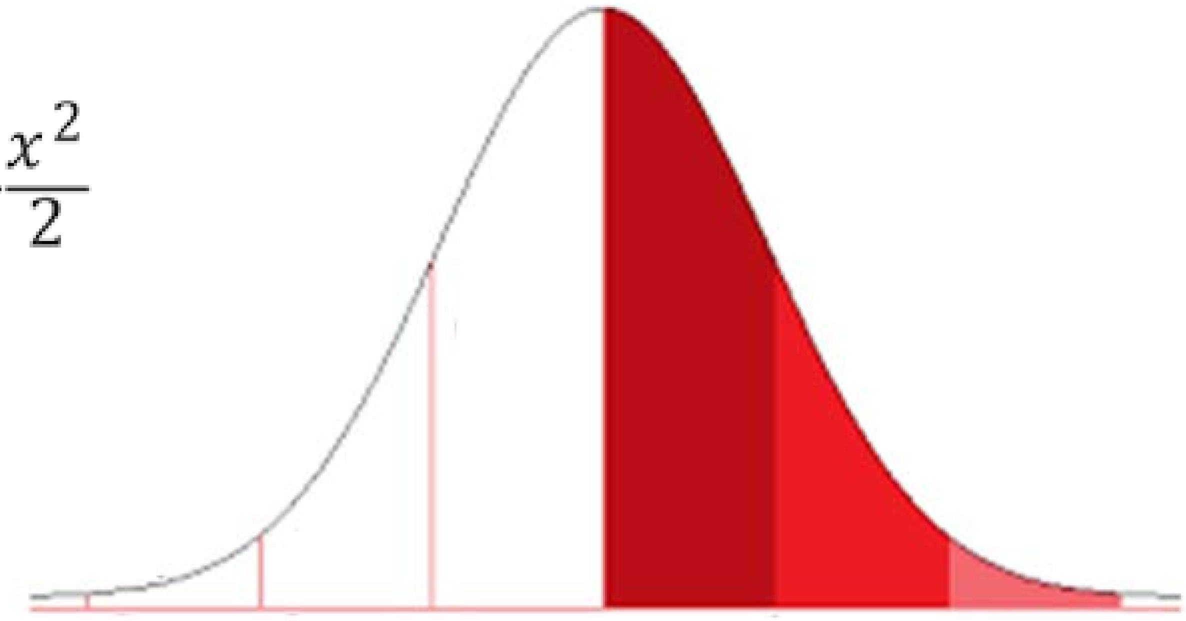




美丽的数学

You are more beautiful than normal curve !

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



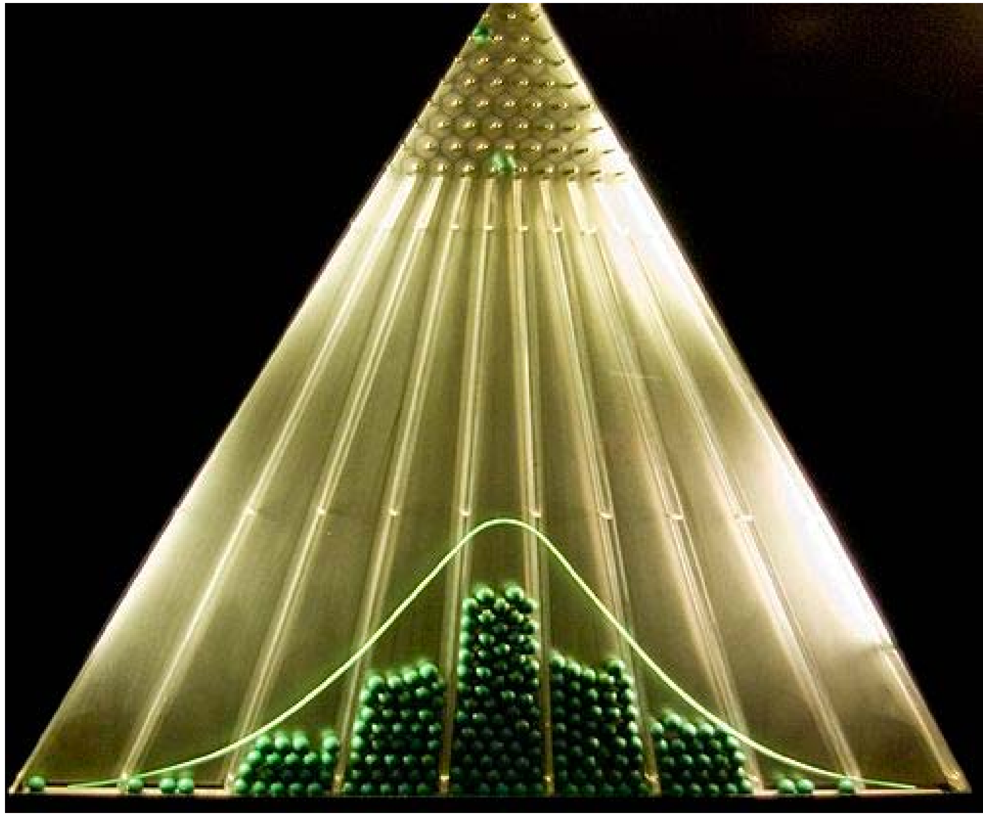
远古的人类为什么要保存火种 ……

夜 幕 来 临 之 际
火 光 摇 曳 妩 媚
灿 烂 多 姿
是 最 美 最 美 的……

Thanks for your attentions！

正态魅影

- 二项分布 $B(n,p)$ 在 n 很大逼近正态分布 $N(np, np(1-p))$
- 泊松分布 $Poisson(\lambda)$ 在 λ 较大时逼近正态分布 $N(\lambda, \lambda)$
- $\chi^2_{(n)}$ 在 n 很大的时候逼近正态分布 $N(n, 2n)$
- t 分布在 n 很大时逼近标准正态分布 $N(0,1)$
- 正态分布的共轭分布还是正态分布



正态魅影

- 几乎所有的极大似然估计在样本量 n 增大的时候都趋近于正态分布
- 克拉美分解定理 (之前介绍过)：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
- 如果 X, Y 独立且满足正态分布 $N(\mu, \sigma^2)$, 那么 $X + Y$, $X - Y$ 独立且同分布，而正态分布是唯一满足这一性质的概率分布
- 对于两个正态分布 X, Y, 如果 X, Y 不相关则意味着 X, Y 独立，而正态分布是唯一满足这一性质的概率分布