最小二乘法与正态分布

从高斯谈起

南京大学 生命科学学院

2018 年 11 月 1 日

There are three kinds of lies: lies, damned lies, and statistics.

卡尔·弗里德里希·高斯 (Carl Friedrich Gauß)

- 数学家、物理学家、天文学家
- 有“数学王子”之称
- 以“高斯”命名的成果达 110 个

- 1766 年，约翰·提丢斯 (Johann Daniel Titius) 发现行星到太阳距的距离比例有一定联系
- 1801 年 1 月，朱塞普·皮亚齐 (Giuseppe Piazzi) 观测到一颗“星星”
- 同年，朱塞普·皮亚齐病倒，痊愈后“星星”已无法找到

Theorem

提丢斯 -波得定则 假设地球与太阳的距离为 *1*，行星离太阳距离 $d = 0.4 + 0.3 * 2^{n-2}$ ，n 为行星的序号。

- 1801 年 9 月，天文学家的争执引起高斯的注意，并决定通过计算找出
- 1801 年 12 月，高斯宣布计算出行星位置，并得到观测实验的证实
- 1809 年，高斯在其著作《天体运动论》中公布方法——最小二乘法 (Least squares method)

最小二乘法

- 误差的判据：观测值与理论值差的平方和
- 目标：找出最匹配的曲线 $y = \beta_1 + \beta_2 x$
- 理想曲线：使所有观察值的残差平方和达到最小

- 高斯认为
  - 回归分析的最小二乘法是最优的
- 高斯希望找到满足最小二乘法的误差密度函数

Theorem

Gauss-Markov 定理 在给定经典线性回归的假定下，最小二乘估计量是具有最小方差的线性无偏估计量.

设真实值为 θ,$x_1, x_2, \ldots, x_n$ 为 n 次相互独立的测量值，每次测量的误差为 e_i，假设误差的密度函数为 $f(e)$，则测量值的联合概率为 n 个误差的联合概率，为

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(e_i) = \prod_{i=1}^{n} f(x_i - \theta)$$

高斯认为取 $L(\theta)$ 最大值可作为 θ 的估计值，即 $L(\theta)$ 称为样本的似然函数，得到的估计值称为极大似然估计

- 高斯猜测：极大似然估计 = 算术平均值
  - 千百年来大家都认为算术平均是一个好的估计，那极大似然估计导出的就应该是算术平均！

求极大似然估计，令

$$\frac{d \ln L(\theta)}{d\theta} = \frac{d}{d\theta} \sum_{i=1}^{n} \ln f(x_i - \theta) = 0$$

整理得

$$\sum_{i=1}^{n} \frac{f'(x_i-\theta)}{f(x_i-\theta} = 0$$

使用 $g(x) = \frac{f'(x)}{f(x)}$ 替换，得到

$$\sum_{i=1}^{n} g(x_i-\theta) = 0$$

假定 $n = 2$

$$g(x_1-\bar{x}) + g(x_2-\bar{x})$$

且 x_1,x_2 任意，则有

$$g(-x) = -g(x)$$

再令 $n = m+1$，且 $x_1 = x_2 = \cdots = x_m = -x, x_{m+1} = mx$

$$\sum_{i=1}^{n} g(x_i - \bar{x}) = mg(-x) + g(mx)$$

$$\therefore g(mx) = mg(x)$$

唯一满足的连续函数：$g(x) = cx$

概率密度函数???

对其积分，得：

$$\ln y = \frac{1}{2}cx^2 + C$$

$$\Rightarrow f(x) = Me^{cx^2}$$

正态分布

概率密度函数

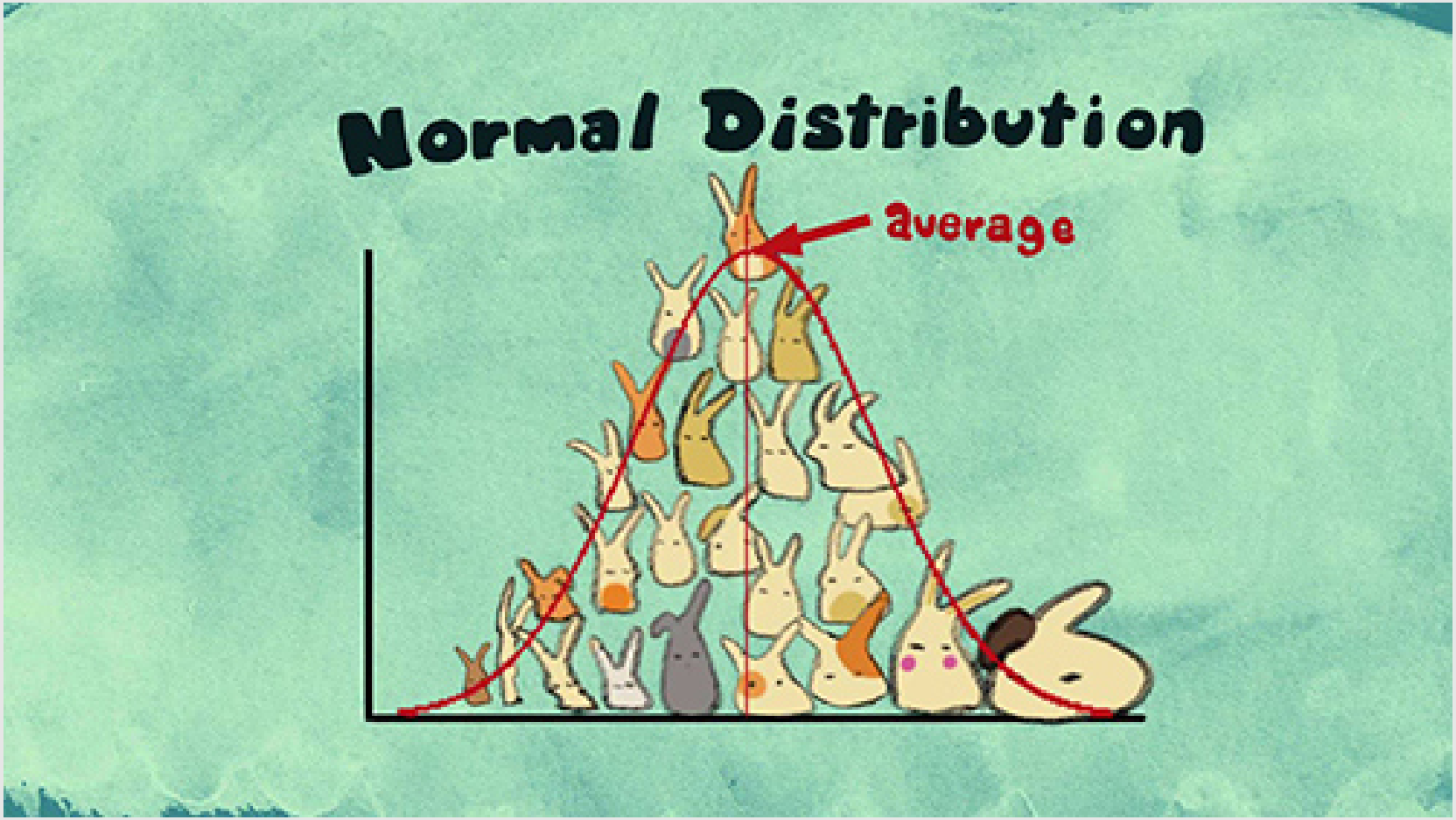
$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

对其进行标准化，使得 $\mu = 1, \sigma = 1$

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

Many years ago I called the Laplace-Gaussian curve the normal curve, which name,while it avoids an international question of priority, has the disadvantage ofleading people to believe that all other distributions of frequency are in onesense or another "abnormal".

-Karl Pearson (1920)


Normal Distribution
average

中心极限定理

Theorem

在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布

- 德莫佛 -拉普拉斯定理
- 林德伯格 -列维定理
- 林德伯格 -费勒定理