教书育人、不忘初心，《数理统计》本科课程
— 课程思政 20 讲

田国梁　统计学教授

南方科技大学 • 统计与数据科学系

Email: tiangl@sustech.edu.cn

中国 • 广东 • 深圳

2024 年 04 月 29 日

提纲 (Outline) Part I

第 1 讲 **South**与**Southern**之区别

第 2 讲 用**14 年创新编写**《数理统计》英文教材

第 3 讲 **Bayes**如何译成中文名？英文名如何读？

第 4 讲 全概率公式 (**Law of Total Probability**) 和 Bayes 公式

第 5 讲 自然常数 (**Natural Constant**) $e = 2.718282\cdots$ 的起源

第 6 讲 从矩母函数与密度函数的关系出发，深度理解**国王函数** e^x

第 7 讲 从对数似然函数出发，深度理解**王后函数** $\log(x)$

第 8 讲 **标准正态分布密度**和**蛇吞象公式**

第 9 讲 从函数的**一阶泰勒展开式**到**线段中/外**任何一点之数学表达式

第 10 讲 函数的一阶泰勒展开之**四种形式**

提纲 (Outline) Part II

第 11 讲 **指数分布**与**几何分布**的**无记忆性**

第 12 讲 **二项分布**的**生存函数**与**贝塔分布**的**累积分布函数**之恒等式

第 13 讲 深度理解**中心极限定理**

第 14 讲 **二项分布**的**正态近似**和**泊松近似**

第 15 讲 从**矩估计量**到**Monte Carlo 积分**

第 16 讲 从**KL 散度**的角度来理解**极大似然估计**之定义

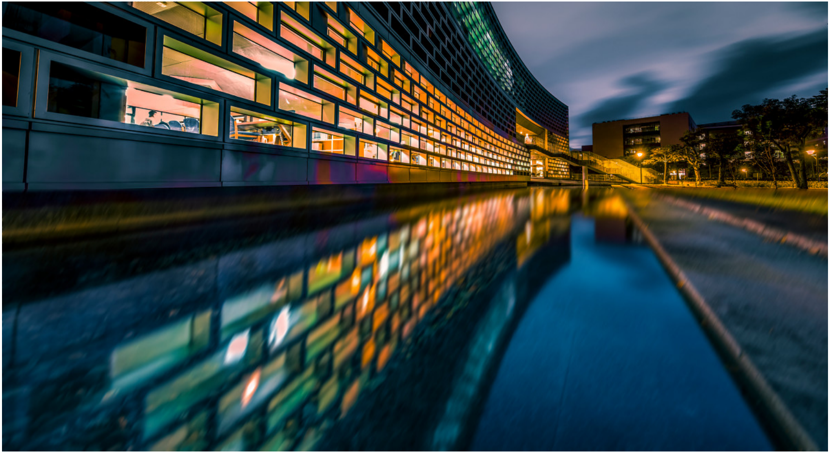
第 17 讲 从**Laplace**提出问题到**Gauss**解决问题：正态分布的发现过程

第 18 讲 度量**点估计量好坏**的指标：**均方误差**

第 19 讲 克拉默–拉奥 (**Cramér–Rao**) 不等式

第 20 讲 建立参数的置信区间过程中的**枢轴量**(**Pivotal Quantity**)

第 16 讲 从 KL 散度的角度来理解极大似然估计之定义



华灯映照图书馆，学子浸醉智海中

16.1 KL 散度 (Kullback–Leibler Divergence)

1° KL 散度之连续版本

- 在概率论和信息论中，KL 散度 (也称为信息散度、相对熵) 是一个**非对称**的指标，以度量**两个概率密度函数 (或 pmf) 之间的差异**。
- 设 $g(\cdot)$ 和 $h(\cdot)$ 是两个具有相同支持 (Support) 的概率密度函数，即:

$$\mathbb{X} = \{x\colon\ g(x) > 0\} = \{x\colon\ h(x) > 0\}.$$

- 设 $X \sim g(x)$，则 g 和 h 之间的 KL 偏差定义为:

$$\mathbf{KL}(g\|h) = E\left\{\log\left[\frac{g(X)}{h(X)}\right]\right\} = \int_{\mathbb{X}} g(x)\log\left[\frac{g(x)}{h(x)}\right]\,\mathrm{d}x \tag{16.1}$$

$$= \int g(x)\log[g(x)]\,\mathrm{d}x - \int g(x)\log[h(x)]\,\mathrm{d}x. \tag{16.2}$$

2° KL 散度关于 g 与 h 是不对称的

- $\mathbf{KL}(g\|h) \neq \mathbf{KL}(h\|g)$。

3° KL 散度之离散版本

- 对于离散情况，设 $\boldsymbol{p} = (p_1, \ldots, p_n)^\top \in \mathbb{T}_n$ 和 $\boldsymbol{q} = (q_1, \ldots, q_n)^\top \in \mathbb{T}_n$ 表示**两个概率向量**，其中 $p_i = \Pr(X = x_i)$, $q_i = \Pr(Y = x_i)$, $i = 1, \ldots, n$, $\mathbb{T}_n \triangleq \mathbb{T}_n(1)$ 且

$$\mathbb{T}_n(c) \triangleq \left\{(p_1, \ldots, p_n)^\top\text{:}\ \ p_i > 0,\ i = 1, \ldots, n,\ \textstyle\sum_{i=1}^n p_i = c\right\}, \tag{16.3}$$

这里 c 是一个正常数。

- 度量 $\boldsymbol{p}$ 与 $\boldsymbol{q}$ 之间的差异之 KL 散度定义为

$$\mathbf{KL}(\boldsymbol{p}\|\boldsymbol{q}) \overset{(16.2)}{=} \sum_{i=1}^n p_i \log(p_i) - \sum_{i=1}^n p_i \log(q_i), \tag{16.4}$$

4° KL 散度是非负的

- $\mathbf{KL}(g\|h) \geqslant 0$, 且等号成立的充要条件是 $g(x) = h(x)$。

5° Jensen 不等式

◆ 设 $\varphi(\cdot)$ 是一个凸 (Convex) 函数。如果随机变量 Y 的取值范围等于 $\varphi(\cdot)$ 的定义域，则

$$E[\varphi(Y)] \geqslant \varphi[E(Y)], \tag{16.5}$$

只要 $E(Y)$ 和 $E[\varphi(Y)]$ 均存在。

6° 证明：$\mathbf{KL}(g\|h) \geqslant 0$

- 在 (16.5) 中，取 $\varphi(\cdot) = -\log(\cdot)$，我们有

$$\begin{aligned}
\mathbf{KL}(g\|h) &\overset{(16.1)}{=} E\left\{\log\left[\frac{g(X)}{h(X)}\right]\right\} = E\left\{-\log\left[\frac{h(X)}{g(X)}\right]\right\} \\
&\overset{(16.5)}{\geqslant} -\log\left\{E\left[\frac{h(X)}{g(X)}\right]\right\} = -\log\left\{\int \frac{h(x)}{g(x)} \cdot g(x)\,\mathrm{d}x\right\} \\
&= -\log 1 = 0.
\end{aligned}$$

$7^\bullet$ 参数向量 θ 的 MLE 之定义

- 设 $X_1, \ldots, X_N \overset{\text{iid}}{\sim} g(x)$, 其中 $g(x)$ 表示母体随机变量 X 的真正的概率密度函数 (True pdf)。
- 既然 $g(x)$ 是永远未知的, 我们想在如下的**假设的密度函数族**

$$\{f(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

中找到一个成员 $f(x;\hat{\boldsymbol{\theta}})$, 作为 $g(x)$ 的最佳近似, 其中 $\hat{\boldsymbol{\theta}}$ 表示 $\boldsymbol{\theta}$ 的 MLE, 定义为:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\Theta}} \sum_{i=1}^{N} \log[f(X_i;\boldsymbol{\theta})] = \arg\max_{\boldsymbol{\Theta}} \frac{1}{N}\sum_{i=1}^{N} \log[f(X_i;\boldsymbol{\theta})]. \tag{16.6}$$

8° Monte Carlo 积分

- 设$X_1,\ldots,X_N \overset{\text{iid}}{\sim} g(x)$且 $X\sim g(x)$, 用样本均值估计母体均值

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to E(X) = \int x g(x)\,\mathrm{d}x, \quad \textbf{as} \quad N\to\infty. \tag{16.7}$$

- 定义 $Y_i = \log[f(X_i;\boldsymbol{\theta})]$, $i=1,\ldots,N$, 且 $Y=\log[f(X;\boldsymbol{\theta})]$, 则

$$\frac{1}{N}\sum_{i=1}^{N}\log[f(X_i;\boldsymbol{\theta})] = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad \textbf{as} \quad N\to\infty$$

$$\overset{(16.7)}{\to} E(Y) = E\Big\{\log[f(X;\boldsymbol{\theta})]\Big\} = \int g(x)\log[f(x;\boldsymbol{\theta})]\,\mathrm{d}x,$$

$$\overset{(16.2)}{=} \underbrace{\int g(x)\log[g(x)]\,\mathrm{d}x}_{\textbf{free from } \boldsymbol{\theta}} - \mathbf{KL}[g(x)\|f(x;\boldsymbol{\theta})]. \tag{16.8}$$

9° MLE 与 KL 散度的联系

- 关于 $\boldsymbol{\theta}$, 通过极大化 (16.8) 的两边, 我们得到:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\Theta}} \ \mathbf{KL}\Big[\underbrace{g(x)}_{\text{true pdf}} \ \Big\| \ \underbrace{f(x;\boldsymbol{\theta})}_{\text{assumed pdf family}}\Big], \quad N \to \infty. \tag{16.9}$$

10° 深度理解 KL 散度的定义

- g 和 h 是两条正的曲线, 其比例的对数 $\log[g(x)/h(x)]$ 是月球上的函数, 代表 $g(x)$ 和 $h(x)$ 在月球上的差异: $\log[g(x)] - \log[h(x)]$。
- 再对 $g(x)$ 积分, 则 (16.1) 表明: KL 散度是一个太空中的函数, 而 g 和 h 这两条曲线在太空 (**希尔伯特空间**) 中, 是两个点。
- $g(x)$ 是一个客观存在且不动的点, $\mathrm{KL}(g(x)\|h(x))$ 散度是度量 $g(x)$ 与 $h(x)$ 这两个点在太空中的**"距离"**。

16.3 所包含的思政元素 (Part I)

- 为了理解 KL 散度的定义 (16.1), 我们可以将客观存在且唯一的 g, 当作**毛主席**, 将变化的不唯一的 h, 当作扮演毛主席的演员的集合, 例如**唐国强、古月、张铁林、刘烨、黄海冰、王仁君、谷智鑫、侯京健、李易峰、佟瑞欣**。让 $\mathrm{KL}(g\|h)$ 达到最小, 等价于在所有扮演毛主席的演员中, 找到一位既形似又神似的一个演员。注意: g 与 h 不能交换位置。

- 而 (16.9) 表明, 极小化 $\mathbf{KL}[g(x)\|f(x;\boldsymbol{\theta})]$, 即对客观存在且唯一的真 pdf $g(x)$, 在实际当中, 我们猜想 $X_1,\ldots,X_N \overset{\text{iid}}{\sim} f(x;\boldsymbol{\theta})$, 然后从密度族 $\{f(x;\boldsymbol{\theta})\colon\ \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ 中找到一个成员, $f(x;\hat{\boldsymbol{\theta}})$, 使得它与 $g(x)$ 的 KL"距离" 达到最短。而该 $\hat{\boldsymbol{\theta}}$ 就是 $\boldsymbol{\theta}$ 的 MLE, 且 $\hat{\boldsymbol{\theta}}$ 依赖于你的猜想 f。

16.4 所包含的思政元素 (Part II)

- 利用**逆向思维**, 将 Monte Carlo 积分公式反向使用, 就得到了公式 (16.8), 使得 n 个独立同分布同参数的随机变量 $\log[f(X_i;\theta)]$ 之平均值与 KL 散度联系起来了。

- **KL 散度**在社会生活和自然科学中客观存在、有用、有意义。它有机地, 将概率论、信息论和统计学紧密地联系在一起。