概 率 论 与 数 理 统 计

Probability and Statistics

南京大学 张绍群 & 徐科

更新： December 2, 2025

Part X – Ch10: 参数估计

Ch10: 参数估计

Parameter Estimation

December 2, 2025

引言

假设我们已知南京大学男性学生的身高服从正态分布 $\mathcal{N}(\mu, \sigma^2)$, 但不知道参数 μ 和 σ 具体的取值. 一种估计参数的方法

- 抽样多个个体, 作为样本
- 通过样本计算样本均值
- 利用样本均值来推断总体的均值 μ

这类已知总体分布形式, 但不知其具体参数, 用样本统计量来估计总体的参数的问题称为 **参数估计问题**. 参数估计是统计推断的核心问题之一, 方法大体上有两类:

- 点估计
- 区间估计

提纲：点估计

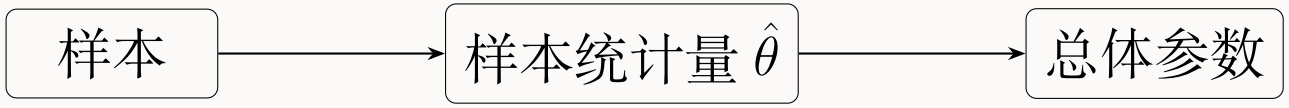
- 矩估计
- 最大似然估计
- 估计量的评价标准
  - 无偏性
  - 有效性：Rao-Crammer 不等式
  - 一致性
- 点估计与机器学习

点估计

定义 0.1 设 $X_1, X_2, \ldots, X_n$ 是来自总体的一个样本, 用于估计未知参数 θ 的统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ 称为 θ 的估计量, 或称为 θ 的 点估计.

Remarks: 点估计的本质就是用样本统计量直接作为总体参数的估计值

- 这里的参数是总体的属性, 而统计量是针对样本的计算.
- 样本通常来自于观测 (数据) 或者采样.
- 点估计流程:



  - 在这里如何构造 $\hat{\theta}$ 没有明确的规定,
  - 1900 年 K. 皮尔逊提出一个替换原理, 即 **矩估计法**.
  - 1922 年费希尔提出的最大似然法, 即 **最大似然估计**.

矩估计法

替换原理具体为:

- 用样本矩去替换总体矩 (这里的矩可以是原点矩也可以是中心距).
  - 使用原点矩
    - 总体 X 的 k 阶原点矩: $a_k = \mathbb{E}[X^k]$
    - 样本的 k 阶原点矩: $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
  - 使用中心矩
    - 总体 X 的 k 阶中心矩: $b_k = \mathbb{E}[(X - \mathbb{E}(X))^k]$
    - 样本的 k 阶中心矩: $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$
- 用样本矩的函数去替换相应的总体矩的函数.

矩估计法 – – 适用场景

根据这个替换原理, 在总体分布形式未知的情况下也可以对参数做出估计, 譬如:

- 用样本均值 $\bar{X}$ 估计总体均值 $\mathbb{E}(X)$
- 用样本方差 S^2 估计总体方差 $\mathbb{VAR}(X)$
  - 注意: 若没有特殊说明, 样本方差采用无偏方差
- 用事件 A 出现的频率估计事件 A 发生的概率.

矩估计法 – – 计算步骤

总体 X 的分布函数 F 包含 m 个未知参数 $\theta_1, \theta_2, \ldots, \theta_m$

- 计算总体 X 的 k 阶矩: $a_k = a_k(\theta_1, \theta_2, \ldots, \theta_m) = \mathbb{E}[X^k], k \in [m]$
  - a_k 一般为 $\theta_1, \theta_2, \ldots, \theta_m$ 的函数
- 计算样本的 k 阶矩: $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
- 令样本矩等于总体矩:

$$A_k = a_k = a_k(\theta_1, \theta_2, \ldots, \theta_m), \quad k = [m]$$

  得到 m 个关于 $\theta_1, \theta_2, \ldots, \theta_m$ 的方程组
- 求解方程组得到估计量 $\hat{\theta_1}, \hat{\theta_2}, \ldots, \hat{\theta_m}$

矩估计：例 0.1

例 0.1 设总体 X 的概率密度函数为

$$f(X)=\begin{cases}(\alpha+1)X^{\alpha}, & X\in(0,1)\\ 0, & \text{其它}\end{cases}$$

设 $X_1,X_2,\ldots,X_n$ 是来自总体的样本, 求参数 α 的矩估计.

解答：例 0.1

题目: 设总体 X 的概率密度函数为

$$f(X)=\begin{cases}(\alpha+1)X^{\alpha}, & X\in(0,1)\\ 0, & \text{其它}\end{cases}$$

设 $X_1,X_2,\ldots,X_n$ 是来自总体的样本, 求参数 α 的矩估计.

解答:

- 首先计算总体 X 的 1 阶矩:

$$\mathbb{E}[X]=\int_{-\infty}^{+\infty}Xf(X)\,\mathrm{d}X=\int_0^1 X(\alpha+1)X^{\alpha}\,\mathrm{d}X=\frac{\alpha+1}{\alpha+2}.$$

  以及样本的 1 阶矩: $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$.
- 根据矩估计方法有

$$\mathbb{E}[X]=\frac{\alpha+1}{\alpha+2}=\bar{X}$$

  求解可得 $\alpha=(2\bar{X}-1)/(1-\bar{X})$.

矩估计：例 0.2

例 0.2 设 $X_1, X_2, \ldots, X_n$ 是来自总体的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{X-\mu}{\theta}}\ , & X \geq \mu\,; \\ 0\,, & \text{其它}\ , \end{cases}$$

其中 $\theta > 0$, 求参数 μ 和 θ 的矩估计.

解答：例 0.2

题目: 如上所述.

解答:

- 设随机变量 $Y = X - \mu$, 则 Y 服从参数为 $1/\theta$ 的指数分布, 有

$$\mathbb{E}[Y] = \theta \quad \text{和} \quad \mathbb{VAR}[Y] = \theta^2.$$

  由此可得 $\mathbb{E}[X] = \mu + \theta$ 和 $\sigma(X) = \theta^2$.
- 计算对应的样本矩

$$A_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad B_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

  求解方程组

$$\mu + \theta = A_1 \quad \text{和} \quad \theta^2 = B_2.$$

  可得 $\mu = \bar{X} - \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}$ 和 $\theta = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$.

最大似然估计法 – – 引子

为了叙述最大似然估计的直观想法, 先看下面这个例子:

例 0.3 设有两个外形相同的箱子中各有 100 只球, 其中甲箱中有 99 个白球、1 个黑球, 乙箱中有 1 个白球、99 个黑球. 今随机抽取一箱并从中抽取一球, 结果取得白球, 问这个白球是从哪个箱子中取出?

最大似然估计法 – – 引子

为了叙述最大似然估计的直观想法, 先看下面这个例子:

例 0.4 设有两个外形相同的箱子中各有 100 只球, 其中甲箱中有 99 个白球、1 个黑球, 乙箱中有 1 个白球、99 个黑球. 今随机抽取一箱并从中抽取一球, 结果取得白球, 问这个白球是从哪个箱子中取出?

解答: A 表示事件 “从甲箱中取出白球”, B 表示事件 “从乙箱中取出白球”, 又有

$$P(A) = 0.99 > P(B) = 0.01$$

因此, 按照可以推断白球 “最可能” 是从甲箱中取出的.

这个推断符合人们的经验事实, 这里的 “最可能” 就是 “极大似然” 之意, 这种想法常称为 “极大似然原理”. 即, 已经得到了样本, 然后通过样本倒推, 找到能够使的该样本出现的最大概率的条件.

最大似然估计法

定义 0.2 设总体的概率函数为 $p(X;\theta)$, $\theta \in \Theta$, 其中 θ 是一个未知参数或几个未知参数组成的参数向量, Θ 是参数空间. $X_1, X_2, \ldots, X_n$ 是来自总体的样本 (默认 i.i.d), 将样本的联合概率函数看成 θ 的函数, 用 $L(\theta, X_1, X_2, \ldots, X_n)$ 表示, 简记 $L(\theta)$,

$$L(\theta) = L(\theta, X_1, X_2, \ldots, X_n) = p(X_1;\theta)p(X_2;\theta)\ldots p(X_n;\theta) \ ,$$

$L(\theta)$ 称为样本的似然函数. 若某个统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ 满足,

$$L(\hat{\theta}) = \max_{\theta\in\Theta} \ L(\theta) \ ,$$

则称 $\hat{\theta}$ 是 θ 的 最大似然估计, 简记为 MLE (Maximum Likelihood Estimation).

最大似然估计法 – – 计算步骤

求 $L(\theta) = p(X_1;\theta)p(X_2;\theta)\ldots p(X_n;\theta)$ 的最大值可以通过下列步骤:

① 列出 $L(\theta) = p(X_1;\theta)p(X_2;\theta)\ldots p(X_n;\theta)$

↓

② 对等式两边取对数 $\log L(\theta) = \sum_{i=1}^{n} \log p(X_i;\theta)$

↓

③ 求关于 θ 一阶偏导并令其为零 $\frac{\partial}{\partial\theta}\log L(\theta) = 0$

↓

④ 求解方程组得到最大似然估计 $\hat{\theta} = \arg\max_\theta L(\theta)$

如何构造似然函数? 核心: 条件概率公式!

最大似然估计：例 0.5

例 0.5 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim Ber(1, p)$ 的样本, 求参数 p 的最大似然估计.

解答：例 0.5

题目: 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim Ber(1, p)$ 的样本, 求参数 p 的最大似然估计.

解答:

- 首先计算似然函数

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i},$$

由而可得对数似然函数

$$\ln L(p) = \sum_{i=1}^{n} X_i \ln p + \left(n - \sum_{i=1}^{n} X_i\right) \ln(1-p),$$

求一阶偏导并令其为零, 可得

$$\frac{\partial \ln L(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right) = 0.$$

由此求解 $p = \sum_{i=1}^{n} X_i/n = \bar{X}$.

最大似然估计：例 0.6

例 0.6 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim U(0, \theta)$ 的样本, 求参数 θ 的最大似然估计.

解答：例 0.6

题目: 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim U(0, \theta)$ 的样本, 求参数 θ 的最大似然估计.

解答:

- 首先计算似然函数

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{I}_{\{0<X_i\leq\theta\}} = \frac{1}{\theta^n} \mathbb{I}_{\{0<X_i\leq\theta\}} \ ,$$

  要使 $L(\theta)$ 最大, 首先是示性函数取值应该为 1, 其次是 $1/\theta^n$ 尽可能大, 由于 $1/\theta^n$ 是 θ 的单调递减函数, 所以 θ 的取值应尽可能小, 但示性函数为 1 决定了 θ 不能小于 $X_{(n)}$, 由此给出 θ 的最大似然估计为 $X_{(n)}$.

- 这个例子说明虽然求导函数是求最大似然估计最常用的方法, 但并不是所有场合求导都是有效的.
- 如果是 $X \sim U(\alpha, \theta)$ 呢?

最大似然估计 – – 不可变性

最大似然估计有一个简单而有效的性质:

定理 0.1 如果 $\hat{\theta}$ 是参数 θ 的最大似然估计, 那么对于任一的函数 $g(\cdot)$, $g(\hat{\theta})$ 也是 $g(\theta)$ 的最大似然估计.

该性质称为最大似然估计的不变性, 从而使得一些复杂结构的参数的最大似然估计的计算变得容易了.

最大似然估计：例 0.7

例 0.7 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的样本, 求参数 μ 和 $\sigma > 0$ 的最大似然估计.

解答：例 0.7

解答:

- 根据正态分布的密度函数, 可知似然函数

$$L(\mu,\sigma)=\frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\left(-\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}\right).$$

  其对数似然函数为 $\ln L(\mu,\sigma)=-n\ln(2\pi)^{1/2}-n\ln\sigma-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2$.

- 对参数 μ 求导计算可得

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\mu}=\sum_{i=1}^{n}(X_i-\mu)=0\Rightarrow\mu=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X},$$

  对参数 σ 求导计算, 可得

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2=0\Rightarrow\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

  根据最大似然估计的不变性, 可知方差 σ^2 的最大似然估计为 $\sigma^2=\sum_{i=1}^{n}(X_i-\bar{X})^2/n$.

最大似然估计：例 0.8

例 0.8 设 $X_1, X_2, \ldots, X_n$ 是来自总体的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \theta \ \mathrm{e}^{-(X-\mu)\theta}, & X \geq \mu\,; \\ 0\,, & \text{其它}\ . \end{cases}$$

求参数 μ 和 θ 的最大似然估计.

解答：例 0.8

题目: 设 $X_1, X_2, \ldots, X_n$ 是来自总体的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \theta \ \mathrm{e}^{-(X-\mu)\theta}, & X \geq \mu ; \\ 0, & \text{其它} . \end{cases}$$

求参数 μ 和 θ 的最大似然估计.

解答:

- 列出似然函数

$$L(\theta, \mu) = \begin{cases} \theta^n \, \mathrm{e}^{-\theta \sum_{i=1}^{n}(X_i - \mu)}, & X_i \geq \mu \\ 0, & \text{其它} \end{cases}$$

其对数似然函数为

$$\ln L(\theta, \mu) = n \ln \theta - \theta \sum_{i=1}^{n} (X_i - \mu).$$

- 对参数 θ 求导计算可得

$$\frac{\partial \ln L(\theta,\mu)}{\partial \theta}=\frac{n}{\theta}-\sum_{i=1}^{n}(X_i-\mu)=0\Rightarrow\theta=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)},$$

对参数 μ 求导计算可得

$$\frac{\partial \ln L(\theta,\mu)}{\partial \mu}=n\theta=0\Rightarrow\theta=0.$$

此时无法求解 μ 和 θ 的最大似然估计.

- 回顾似然函数

$$L(\theta,\mu)=\begin{cases}\theta^n\,\mathrm{e}^{-\theta\sum_{i=1}^{n}(X_i-\mu)}, & X_i\geq\mu\\ 0, & \text{其它}\end{cases}$$

可以发现 μ 越大似然函数 $L(\theta,\mu)$ 越大, 但须满足 $X_i\geq\mu(i\in[n])$. 由此可得最大似然估计为

$$\hat{\mu}=X_{(1)}\quad\text{and}\quad\hat{\theta}=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}(X_i-X_{(1)})}\,.$$

最大似然估计：例 0.9

例 0.9 设总体 X 的概率密度函数为

$$f(X)=\begin{cases}(\alpha+1)X^{\alpha}, & X\in(0,1)\\ 0, & \text{其它}\end{cases}$$

设 $X_1,X_2,\ldots,X_n$ 是来自总体的样本, 求参数 α 的矩估计.

解答：例 0.9

题目: 如上所述.

解答:

- 列出似然函数

$$L(\alpha) = (\alpha+1)^n \prod_{i=1}^{n} X_i^{\alpha} = (\alpha+1)^n (X_1 X_2 \dots X_n)^{\alpha},$$

以及其对数似然函数为 $\ln L(\alpha) = n\ln(\alpha+1) + \alpha \ln(X_1 X_2 \dots X_n)$. 求导并令导数为零有

$$\frac{\partial \ln L(\alpha)}{\partial \alpha} = \frac{n}{\alpha+1} + \ln(X_1 X_2 \dots X_n) = 0,$$

求解可得

$$\alpha = \frac{-n}{\sum_{i=1}^{n} \ln(X_i)} - 1$$

对比例 0.1, 可以看到同一密度函数的矩估计和最大似然估计结果可能不同.

估计量的评价标准

不同的估计方法可能得到不同的估计值.

自然地, 我们希望知道采用哪一种估计量更好, 或更好的标准是什么呢? 统计学上, 给出了无偏性、有效性、一致性等评价标准.

- 无偏性: $\hat{\theta}$ 与参数真值 θ 之间的偏差的平均值为 0
- 有效性: $\hat{\theta}$ 围绕参数真值 θ 的方差越小越好
- 一致性: 随着样本量的不断增大, $\hat{\theta}$ 能够有效逼近参数真值 θ

无偏性

定义 0.3 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ 是 θ 的一个估计, θ 的参数空间为 Θ, 若对任意的 $\theta \in \Theta$, 有

$$\mathbb{E}_\theta(\hat{\theta}) = \theta,$$

则称 $\hat{\theta}$ 是 θ 的 *无偏估计*, 否则称为 *有偏估计*.

Remarks:

- (原点矩) 样本 k 阶原点矩为总体 k 阶原点矩的无偏估计
- (中心矩) 设 $X_1, X_2, \ldots, X_n$ 来自总体 X 的样本, 期望 μ, 方差 σ^2, 则:
  - $S_0^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的有偏估计
  - $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ 是 σ^2 的无偏估计
- 若 $\hat{\theta}$ 是 θ 的一个无偏估计, $g(\hat{\theta})$ 不一定也是 $g(\theta)$ 的无偏估计.

无偏估计：例 0.10

例 0.10 设 $X_1, X_2, \ldots, X_n$ 来自总体 X 的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{X}{\theta}}\ , & X \geq 0\,; \\ 0\,, & X < 0\,. \end{cases}$$

证明: 统计量

$$\bar{X} = \sum_{i=1}^{n} X_i/n \quad \text{和} \quad n \cdot \min\{X_1, X_2, \ldots, X_n\}$$

均是 θ 的无偏估计.

解答：例 0.10

解答:

- 根据期望和指数分布的性质, 有

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \theta,$$

  由此可知, $\bar{X}$ 是 θ 的无偏估计 (原点矩) .

- 设随机变量 $Z = \min\{X_1, X_2, \ldots, X_n\}$, 则有

$$\begin{aligned} F_Z(z) &= P(Z \le z) = 1 - P(Z > z) \\ &= 1 - P(X_1 > z)P(X_X > z)\ldots P(X_n > z) \\ &= 1 - \prod_{i=1}^{n}(1 - P(X_i \le z)) = \begin{cases} 0, & z < 0 \\ 1 - \mathrm{e}^{-nz/\theta}, & z \ge 0 \end{cases} \end{aligned}$$

  于是当 $z \ge 0$ 时, 有

$$P(Z > z) = 1 - F_Z(z) = \mathrm{e}^{-nz/\theta}.$$

根据指数分布期望的性质, 有

$$\mathbb{E}[Z] = \int_0^{\infty} P(Z > z)\, dz = \int_0^{+\infty} \mathrm{e}^{-nz/\theta}\, \mathrm{d}z = \frac{\theta}{n},$$

于是有 $\theta = \mathbb{E}[nZ]$.

有效性

例子 0.10 说明: 参数可能存在多个无偏估计.

- 若 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计, 如何在多个无偏估计中进行选择?
- 直观的想法是, $\hat{\theta}$ 围绕参数真值 θ 的方差越小越好, 即有效性.

定义 0.4 设 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 分别是 θ 的两个无偏估计, 如果对任意的 $\theta \in \Theta$ 都有

$$\mathbb{VAR}(\hat{\theta}_1) \leq \mathbb{VAR}(\hat{\theta}_2) ,$$

且至少有一个 $\theta \in \Theta$ 使得上述不等式严格成立, 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

Remarks: 有效性是针对无偏估计而言的, 因此判断有效性之前必须先确认估计量的无偏性.

有效性：例 0.11

例 0.11 设 $X_1, X_2, \ldots, X_n$ 来自总体 X 的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{X}{\theta}}\ , & X \geq 0\,; \\ 0\,, & X < 0\,. \end{cases}$$

令 $Z = \min\{X_1, X_2, \ldots, X_n\}$. 证明: 当 $n > 1$ 时, $\bar{X} = \sum_{i=1}^{n} X_i/n$ 比 nZ 更有效.

解答：例 0.11

题目: 如上所述.

解答:

- 根据样本的独立性有

$$\sigma(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma(X_i) = \frac{\theta^2}{n}.$$

又根据例0.10可知随机变量 Z 的密度函数为

$$f(z) = \begin{cases} 0, & z < 0 \\ \frac{n}{\theta}\,\mathrm{e}^{-\frac{nz}{\theta}}, & z \geq 0 \end{cases}$$

从而得到

$$\sigma(nZ) = n^2\sigma(Z) = n^2\frac{\theta^2}{n^2} = \theta^2,$$

因此当 $n \geq 1$ 时有 $\sigma(\bar{X}) \leq \sigma(nZ)$ 成立, 故估计量 $\bar{X}$ 比 nZ 更有效.

有效性：例 0.12

例 0.12 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 且

$$\mathbb{E}(X) = \mu \quad 及 \quad \mathbb{VAR}(X) = \sigma^2 .$$

设常数 $c_1, c_2, \ldots, c_n \geq 0$, 满足 $\sum_{i=1}^{n} c_i = 1, c_i \neq \frac{1}{n}$. 求证: $\bar{X}$ 比 $\sum_{i=1}^{n} c_i X_i$ 有效.

解答：例 0.12

题目: 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 且 $\mathbb{E}(X) = \mu$ 及 $\mathbb{VAR}(X) = \sigma^2$. 设常数 $c_1, c_2, \ldots, c_n \geq 0$, 满足 $\sum_{i=1}^n c_i = 1, c_i \neq \frac{1}{n}$. 求证: $\bar{X}$ 比 $\sum_{i=1}^n c_i X_i$ 有效.

解答:

- 根据样本的独立同分布的性质, 有

$$\mathbb{E}(\bar{X}) = \mu \quad \text{和} \quad \mathbb{VAR}(\bar{X}) = \frac{\sigma^2}{n}.$$

  根据期望的性质, 有 $\mathbb{E}[\sum_{i=1}^n c_i X_i] = \mu$, 进一步有

$$\mathbb{VAR}\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 \mathbb{VAR}(X_i) = \sigma^2 \sum_{i=1}^n c_i^2 \geq \frac{\sigma^2}{n}.$$

  这里利用不等式 $\sum_{i=1}^n c_i^2/n \geq (\sum_{i=1}^n c_i/n)^2$, 所以有 $\mathbb{VAR}(\sum_{i=1}^n c_i X_i) \geq \mathbb{VAR}(\bar{X})$.

Cramér-Rao 不等式

有效性希望 $\hat{\theta}$ 围绕参数真值 θ 的方差越小越好, 那么这个方差能小到什么程度? 有无下界? 若有的话, 如何去求? Cramér-Rao 不等式回答了这些问题.

定理 0.2 随机变量 X 的概率密度为 $f(X;\theta)$ 或概率质量函数为 $p(X;\theta)$, 令

$$\mathbb{VAR}_0(\theta)=\frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]} \quad 或 \quad \mathbb{VAR}_0(\theta)=\frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln p(X;\theta)}{\partial\theta}\right)^2\right]}$$

对任意的无偏估计量 $\hat{\theta}$, 有

$$\mathbb{VAR}(\hat{\theta})\geq\mathbb{VAR}_0(\theta)\ ,$$

称 $\mathbb{VAR}_0(\theta)$ 为估计量 $\hat{\theta}$ 方差的下界. 当 $\mathbb{VAR}(\hat{\theta})=\mathbb{VAR}_0(\theta)$ 时, 称 $\hat{\theta}$ 为达到方差下界的无偏估计量, 此时 $\hat{\theta}$ 为最有效估计量, 简称 有效估计量.

有效性：例 0.13

例 0.13 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{X}{\theta}}\ , & X > 0\,; \\ 0\,, & X \leq 0\,. \end{cases}$$

证明: θ 的最大似然估计为有效估计量.

解答：例 0.13

题目: 如上所述.

解答:

- 根据定理 0.2, 首先计算 $\sigma_0(\theta)$. 又

$$\ln f(X;\theta) = -\ln\theta - \frac{X}{\theta}, \qquad \frac{\partial \ln f(X;\theta)}{\partial\theta} = -\frac{1}{\theta} + \frac{X}{\theta^2}$$

  所以

$$\mathbb{VAR}_0(\theta) = \frac{1}{n\mathbb{E}\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{n\mathbb{E}\left[\left(-\frac{1}{\theta} + \frac{X}{\theta^2}\right)^2\right]} = \frac{1}{\frac{n}{\theta^4}\mathbb{E}[(X - \mathbb{E}[X])^2]} = \frac{\theta^2}{n}$$

- 计算对数似然函数, 有

$$\ln L(\theta) = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} X_i \Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

  进一步得到最大似然估计 $\hat{\theta}$ 的方差 $\mathbb{VAR}(\hat{\theta}) = \frac{\theta^2}{n}$, 因此 θ 的最大似然估计为有效估计量.

一致性

定义 0.5 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ 是 θ 一个估计量. 当 $n \to \infty$ 时, 有 $\hat{\theta} \xrightarrow{P} \theta$ 成立, 即对任意 $\epsilon > 0$, 有

$$\lim_{n \to 0} P\left[|\hat{\theta} - \theta| > \epsilon\right] = 0 \,,$$

则称 $\hat{\theta}$ 为 θ 的 一致估计量.

Remarks: 一致性被认为是对估计的一个最基本要求.

- 如果一个估计量, 在样本不断增多时都不能有效的靠近被估参数的真实值, 那么这个估计是很值得怀疑的.
- 通常, 不满足一致性的估计都不予考虑.

一致性

在判断或计算参数的一致估计量时, 下述两个定理是很有用的.

定理 0.3 (一致性的充分条件) 设 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ 是 θ 的一个估计量, 若满足以下两个条件:

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta_n}] = \theta\ , \qquad \lim_{n\to\infty} \mathbb{VAR}\left[\hat{\theta_n}\right] = 0$$

则 $\hat{\theta}$ 为 θ 的一致估计量.

定理 0.4 (一致性的函数不变性) 设 $\hat{\theta}_{n1}, \hat{\theta}_{n2}, \ldots, \hat{\theta}_{nk}$ 分别是 $\theta_1, \theta_2, \ldots, \theta_k$ 的一致性估计, $G = g(\theta_1, \theta_2, \ldots, \theta_k)$ 是 $\theta_1, \theta_2, \ldots, \theta_k$ 的连续函数, 则

$$\hat{G} = g\left(\hat{\theta}_{n1}, \hat{\theta}_{n2}, \ldots, \hat{\theta}_{nk}\right)$$

是 G 的一致性估计.

一致性：例 0.14

例 0.14 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 且总体 X 的概率密度函数为

$$f(X) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{X}{\theta}}\ , & X > 0\,; \\ 0\,, & X \leq 0\,. \end{cases}$$

证明: 样本均值 $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 是 θ 的一致估计量.

解答：例 0.14

题目: 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 且总体 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{\theta}\ \mathrm{e}^{-\frac{x}{\theta}}\ , & x > 0\ ; \\ 0\ , & x \leq 0\ . \end{cases}$$

证明: 样本均值 $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 是 θ 的一致估计量.

解答:

- 根据定理 0.3, 有

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}] = \theta \quad \text{和} \quad \lim_{n\to\infty} \mathbb{VAR}[\hat{\theta}] = \lim_{n\to\infty} \frac{\theta^2}{n} = 0\ .$$

  证毕.

一致性：例 0.15

例 0.15 设 $X_1, X_2, \ldots, X_n$ 是取自总体 $X \sim U(0, \theta)$ 的样本. 证明: 参数 θ 的最大似然估计是一致估计量.

解答：例 0.15

- 根据前面的例题, 可知 θ 的最大似然估计是 $\hat{\theta} = x_{(n)}$. 设随机变量 $Z = x_{(n)}$, 则 Z 的分布函数为

$$F_Z(z) = P(Z \le z) = P(x_{(n)} \le z) = \prod_{i=1}^{n} P(x_i \le z) = \begin{cases} 1, & z > \theta \\ (\frac{z}{\theta})^n, & z \in [0, \theta] \\ 0, & z < 0 \end{cases}$$

  由此可得当 $z \in [0, \theta]$ 时随机变量 Z 的密度函数为 $f_Z(z) = nz^{n-1}/\theta^n$.

- 进一步有

$$\mathbb{E}[\hat{\theta}] = \int_0^{\theta} \frac{nz^n}{\theta^n} \, \mathrm{d}z = \frac{n}{n+1}\theta \Rightarrow \lim_{n\to\infty} \mathbb{E}[\hat{\theta}] = \theta$$

  又 $\mathbb{E}[Z^2] = \int_0^{\theta} \frac{nz^{n+1}}{\theta^n} \, \mathrm{d}z = \frac{n}{n+2}\theta^2$, 因此有

$$\mathbb{VAR}[\hat{\theta}] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 = \frac{n}{n+2}\theta^2 - (\frac{n\theta}{n+1})^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 \Rightarrow \lim_{n\to\infty} \mathbb{VAR}[\hat{\theta}] = 0$$

  由此, $\hat{\theta}$ 是 θ 的有偏、但一致估计量.

点估计与机器学习

- 在统计学中, 点估计是用样本估计未知总体参数的一个具体数值.
- 在机器学习中, 训练模型其实就是在用数据估计模型参数的过程. 因此，模型参数的训练结果，就是参数的点估计值.
  - 线性回归 $\rightarrow$ 用最小二乘法估计参数（最大似然估计的特例）
  - 逻辑回归 $\rightarrow$ 用最大似然估计参数（点估计）
  - 神经网络 $\rightarrow$ 通过梯度下降最小化损失 (等价于某种似然下的点估计)

点估计与机器学习: 线性回归的最大似然点估计

假设我们有一个二分类问题，样本数据如下表

样本编号	特征 x_i	标签 y_i
1	2.0	1
2	1.0	0
3	3.0	1
4	0.5	0

考虑简单线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \,, \quad i = 1, 2, \ldots, n$$

其中, β_0 和 β_1 是待估参数, ε_i 为独立同分布的正态误差.

Step 1: 构造似然函数

根据正态分布的概率密度函数，样本的联合似然函数为

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right)$$

对数似然函数为

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Step 2: 最大似然估计

最大化对数似然函数关于 β_0, β_1 的部分，可以得到

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{1}$$

这正是 **最小二乘法** 求解的目标函数.

Step 3: 解的闭式表达

令样本均值为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

则最大似然估计（点估计）为

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

因此，$(\hat{\beta}_0, \hat{\beta}_1)$ 就是线性回归模型参数的最大似然点估计.

点估计与机器学习: 逻辑回归的最大似然点估计

假设我们有一个二分类问题，样本数据如下表

样本编号	特征 x_i	标签 y_i
1	2.0	1
2	1.0	0
3	3.0	1
4	0.5	0

采用如下逻辑回归模型进行二分类

$$P(y=1 \mid x; w, b) = \sigma(wx+b) = \frac{1}{1+\mathrm{e}^{-(wx+b)}}$$

Step 1: 构造似然函数

假设样本独立同分布，则联合似然函数为

$$L(w,b)=\prod_{i=1}^{n}[\sigma(wx_i+b)]^{y_i}[1-\sigma(wx_i+b)]^{1-y_i}$$

对数似然函数为：

$$\ell(w,b)=\sum_{i=1}^{n}\Big[y_i\ln\sigma(wx_i+b)+(1-y_i)\ln(1-\sigma(wx_i+b))\Big]$$

Step 2: 最大似然估计

最大化对数似然函数得到参数的点估计

$$(\hat{w}, \hat{b}) = \arg\max_{w,b} \ell(w, b)$$

其偏导数为

$$\begin{cases} \dfrac{\partial \ell}{\partial w} = \displaystyle\sum_{i=1}^{n} (y_i - \sigma(wx_i + b))x_i \\ \dfrac{\partial \ell}{\partial b} = \displaystyle\sum_{i=1}^{n} (y_i - \sigma(wx_i + b)) \end{cases}$$

通过梯度上升法（或等价的最小化负对数似然）可以求得最优参数，例如

$$\hat{w} = 1.52, \quad \hat{b} = -2.01$$

该 $(\hat{w}, \hat{b})$ 即为逻辑回归模型参数的最大似然点估计.

点估计与机器学习: 神经网络中的点估计示例

考虑一个单隐藏层神经网络的回归问题

$$\hat{y}_i = f(x_i; \Theta), \quad i = 1, 2, \ldots, n$$

其中,

- $x_i \in \mathbb{R}^d$ 为输入特征
- $\hat{y}_i \in \mathbb{R}$ 为预测输出
- $\Theta = \{W_1, b_1, W_2, b_2\}$ 为网络参数, 包括输入层到隐藏层的权重 W_1 和偏置 b_1, 以及隐藏层到输出层的权重 W_2 和偏置 b_2
- 激活函数为 $\sigma(\cdot)$, 隐藏层输出为 $h_i = \sigma(W_1 x_i + b_1)$, 输出层为 $\hat{y}_i = W_2 h_i + b_2$

Step 1: 构造似然函数

假设输出服从独立同分布的正态分布

$$y_i \sim \mathcal{N}(\hat{y}_i, \sigma^2)$$

则样本的联合似然函数为

$$L(\Theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2} \right)$$

对数似然函数为

$$\ell(\Theta, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Step 2: 最大似然估计（点估计）

最大化对数似然函数关于 Θ 的部分, 相当于最小化平方误差损失

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{i=1}^{n} (y_i - f(x_i; \Theta))^2$$

通过梯度下降法或其变种（如 Adam）进行迭代优化, 即可得到参数的最大似然点估计 $\hat{\Theta}$.

例如，在训练一个单隐藏层 5 个神经元的网络后，得到

$$\hat{W}_1 = \begin{bmatrix} 0.12 & -0.34 \\ 0.45 & 0.27 \\ \vdots & \vdots \end{bmatrix}, \quad \hat{b}_1 = \begin{bmatrix} 0.01 \\ -0.05 \\ \vdots \end{bmatrix}, \quad \hat{W}_2, \hat{b}_2, \ldots$$

这些 $\hat{\Theta}$ 就是神经网络参数的点估计.