



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第二讲：参数估计-点估计

康雁飞

数量经济与商务统计系

Outline

- 1 参数估计做什么？
- 2 点估计与估计量的评选标准
- 3 矩估计及其统计性质
- 4 极大似然估计

Outline

- 1 参数估计做什么？
- 2 点估计与估计量的评选标准
- 3 矩估计及其统计性质
- 4 极大似然估计

- 上一章中常用统计量及其抽样分布 \Rightarrow 目的在于对感兴趣的问题进行统计推断。
- **统计推断的任务**：根据样本信息，推断总体的统计规律。
- **统计推断的基本问题**：
 - 1 估计问题：根据样本信息，对总体分布中未知参数进行估计（参数估计：本章节重点内容）
 - 2 假设检验问题（L3 重点内容）

参数估计问题的提出

- 分布中所含的未知参数 θ : 两点分布 $b(1, p)$, 正态分布 $N(\mu, \sigma^2)$ 。
- 分布中所含的未知参数 θ 的函数: 服从正态分布 $N(\mu, \sigma^2)$ 的变量不超过 a 的概率 $P(X \leq a) = \Phi\left(\frac{a-\mu}{\sigma}\right)$ 是未知参数 μ, σ 的函数。
- 分布的各种特征函数是未知参数: 均值、方差等。

- **记号**：常用 θ 表示参数，参数 θ 所有可能取值组成的集合称为参数空间，记为 Θ 。
- **参数估计问题**：根据样本构造适当的统计量对各种未知参数做出估计：
 - 1 点估计（本节主要内容）
 - 2 区间估计（下节主要内容）
- **基本要求**：熟悉掌握两种参数估计形式的基本思想、构造方法、统计性质及其在不同分布情形下的计算、R 实现

Outline

- 1 参数估计做什么？
- 2 点估计与估计量的评选标准
- 3 矩估计及其统计性质
- 4 极大似然估计

点估计 (Point estimator)

点估计

设 X_1, X_2, \dots, X_n 是来自总体 X 的一个容量为 n 的样本, 用来估计未知参数 θ 的统计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ 称为 θ 的估计量或点估计, 简称估计 (estimator)。

- 因为样本 X_1, X_2, \dots, X_n 是随机变量, θ 的点估计 $\hat{\theta}$ 也是一个随机变量, 如 $\hat{\mu} = \bar{X}$ 。
- 若 x_1, x_2, \dots, x_n 表示的是样本的观测值, 则 $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ 是 θ 的估计值 (estimate)。
- **注意:** Estimator 是对参数估计的随机变量; estimate 是该 estimator 的一个具体值。

估计量的评选标准

- **思考**：如何构造统计量 $\hat{\theta}$?
- 对总体的未知参数可用不同方法求得不同的估计量，如何评价估计量的好坏呢？
- 常用的三条标准：
 - 1 无偏性 (Unbiasedness)
 - 2 有效性 (Efficiency)
 - 3 相合性 (Consistency)

无偏性

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

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

则称 $\hat{\theta}$ 是 θ 的无偏估计 (unbiased estimator), 否则称为有偏估计 (biased estimator)。

- 偏差 $\text{bias} = E(\hat{\theta}) - \theta$.
- 无偏性是估计是否合理的一个基本标准。

课堂练习一

1 设总体 X 的 k 阶矩 $\mu_k (k \geq 1)$ 存在, 证明无论总体服从什么分布, k 阶样本矩 $A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ 是 μ_k 的无偏估计量 (X_1, X_2, \dots, X_n 为 X 的一个样本)。

2 总体方差 σ^2 的两个估计:

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 是否为无偏估计?

3 设 X 服从参数为 θ 的指数分布 $f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} I(x > 0)$,

X_1, X_2, \dots, X_n 为 X 的一个样本, 证明 \bar{X} 和

$nX_{(1)} = n \cdot \min(X_1, X_2, \dots, X_n)$ 都是 θ 的无偏估计。

课堂练习一

我们可以证明样本方差 $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ 不是总体方差 σ^2 的无偏估计，因为 $E(S_n^2) = \frac{n-1}{n}\sigma^2$ 。对此，有如下两点说明：

- 1 当样本量趋于无穷时，有 $E(S_n^2) = \frac{n-1}{n}\sigma^2 \rightarrow \sigma^2$ ，我们称 S_n^2 为 σ^2 的**渐近无偏估计** (asymptotically unbiased estimator)。
- 2 若对 S_n^2 作如下修正：

$$S^2 = \frac{nS_n^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

则 S^2 是总体方差的无偏估计。

- 1 对任一总体而言，样本均值是总体均值的无偏估计；当总体 k 阶矩存在时，样本 k 阶原点矩是总体 k 阶原点矩的无偏估计。 k 阶中心矩则不一样。
- 2 无偏性不具有不变性：即若 $\hat{\theta}$ 是 θ 的无偏估计，一般而言， $g(\hat{\theta})$ 不是 $g(\theta)$ 的无偏估计，除非 $g(\theta)$ 是 θ 的线性函数（见例 6.1.2）。

思考：当参数的无偏估计很多时，如何在无偏估计中进行选择？

有效性

设 $\hat{\theta}_1, \hat{\theta}_2$ 是 θ 的两个无偏估计，如果对任意的 $\theta \in \Theta$ ，有

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2),$$

且至少有一个 $\theta \in \Theta$ 使得上述不等号严格成立，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ **有效**。

一致最小方差无偏估计

一致最小方差无偏估计

设 $\hat{\theta}$ 为 θ 的一个无偏估计, 如果对任意一个无偏估计 $\tilde{\theta}$, 对 $\theta \in \Theta$, 都有

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}),$$

则称 $\hat{\theta}$ 为 θ 的一致最小方差无偏估计 (Uniformly Minimum Variance Unbiased Estimator), 简记 UMVUE。

- 设 X_1, X_2, \dots, X_n 是取自某总体的样本, 记总体均值为 μ , 总体方差为 σ^2 , 则 $\hat{\mu}_1 = X_1, \hat{\mu}_2 = \bar{X}$ 都是 μ 的无偏估计, 但

$$\text{Var}(\hat{\mu}_1) = \sigma^2, \quad \text{Var}(\hat{\mu}_2) = \sigma^2/n.$$

- 显然, 只要 $n > 1$, $\hat{\mu}_2$ 比 $\hat{\mu}_1$ 有效, 这表明用全部数据的平均估计总体均值要比只使用部分数据更有效。

举例

- 设 X_1, \dots, X_n 是来自均匀总体 $U(0, \theta)$ 中的样本, 人们常用最大观测值 $X_{(n)}$ 来估计 θ (这是 θ 的极大似然估计)。
- 由于 $E(X_{(n)}) = \frac{n}{n+1}\theta$ (教材例 6.2.5), 所以 $X_{(n)}$ 不是 θ 的无偏估计, 而是 θ 的渐近无偏估计。经过修正后可以得到 θ 的一个无偏估计:
 $\hat{\theta}_1 = \frac{n+1}{n}X_{(n)}$, 且

$$\begin{aligned}\text{Var}(\hat{\theta}_1) &= \left(\frac{n+1}{n}\right)^2 \text{Var}(X_{(n)}) \\ &= \left(\frac{n+1}{n}\right)^2 \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{\theta^2}{n(n+2)}.\end{aligned}$$

- 由于总体均值为 $\theta/2$ ，可以使用样本均值估计总体均值，于是可得到 θ 的另一个无偏估计 $\hat{\theta}_2 = 2\bar{X}$ （这是 θ 的矩估计），且

$$\text{Var}(\hat{\theta}_2) = 4 \text{Var}(\bar{X}) = \frac{4}{n} \text{Var}(X) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

- 由此，当 $n > 1$ 时， $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

相合性

思考：无偏性和有效性都是在样本容量 n 固定时提出的，一个自然的想法是：随着 n 的增大，一个估计量的值稳定于待估参数的真值。

相合性

设 $\theta \in \Theta$ 为未知参数， $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ 是 θ 的一个估计量， n 是样本容量，若对任何一个 $\epsilon > 0$ ，有

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta}_n - \theta\right| < \epsilon\right) = 1,$$

则称 $\hat{\theta}_n$ 为 θ 的**相合 (或一致) 估计** (consistent estimator)。

相合性：两个定理

- 1 设 $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots, x_n)$ 是 θ 的一个估计量, 若
- $$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta, \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0,$$

则 $\hat{\theta}_n$ 是 θ 的相合估计。

- 2 若 $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk}$ 分别是 $\theta_1, \dots, \theta_k$ 的相合估计, $\eta = g(\theta_1, \dots, \theta_k)$ 是 $\theta_1, \dots, \theta_k$ 的连续函数, 则 $\hat{\eta}_n = g(\hat{\theta}_{n1}, \dots, \hat{\theta}_{nk})$ 是 η 的相合估计。

设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，由大数定律及上述定理可以推出：**矩估计一般都具有相合性**。比如：

- 样本均值 \bar{X} 是总体均值 μ 的相合估计
- 样本方差 S_n^2 是总体方差 σ^2 的相合估计
- 样本方差 S^2 是总体方差 σ^2 的相合估计（参数的相合估计不止一个）
- 样本标准差 S 是总体标准差 σ 的相合估计

随堂练习二

- 1 设 X_1, X_2, \dots, X_n 是来自均匀总体 $U(0, \theta)$ 的样本, 证明 $X_{(n)}$ 是 θ 相合估计。

随堂练习二

2 设一个试验有三种可能的结果，发生的概率分别为

$$p_1 = \theta^2, \quad p_2 = 2\theta(1 - \theta), \quad p_3 = (1 - \theta)^2$$

现做 n 次试验，观测到三种结果发生的次数分别为 n_1, n_2, n_3 ，可以采取频率替换方法估计 θ 。由于可以有三个不同的 θ 的表达式

$$\theta = \sqrt{p_1}, \quad \theta = 1 - \sqrt{p_3}, \quad \theta = p_1 + p_2/2$$

从而可以给出 θ 三种不同的频率替换估计，分别为

$$\hat{\theta}_1 = \sqrt{n_1/n}, \quad \hat{\theta}_2 = 1 - \sqrt{n_3/n}, \quad \hat{\theta}_3 = \frac{n_1}{n} + \frac{n_2}{2n}$$

由大数定律， $n_1/n, n_2/n, n_3/n$ 分别是 p_1, p_2, p_3 的相合估计， $\hat{\theta}_1, \hat{\theta}_2$ 和 $\hat{\theta}_3$ 也都是 θ 的相合估计。

- 1 什么是点估计？
- 2 估计量优良性评选准则：无偏性、有效性、相合性
- 3 理解并熟练掌握三种优良性准则的定义及直观统计含义
- 4 掌握如何验证估计的无偏性、有效性和相合性

思考：如何用数值模拟的方式验证估计是相合的？

教材习题 6.1: 第 1、4、5、6、7 题

Outline

- 1 参数估计做什么？
- 2 点估计与估计量的评选标准
- 3 矩估计及其统计性质
- 4 极大似然估计

矩估计 (Method of Moments, MOM)

- 早期起源 (19 世纪末至 20 世纪初): 最初由数学家卡尔·皮尔逊 (Karl Pearson) 提出, 他在研究分布时发现, 通过矩 (如均值、方差等) 来推测参数是一种简单有效的方式。
- 正式定义和推广 (20 世纪初至中期): 统计学家如卡尔·皮尔逊和罗斯·皮尔森 (R.A. Fisher) 开始系统研究估计问题, 并发展出了矩估计的理论框架。
- 现代应用和扩展 (20 世纪后期至今): 被应用到更多领域, 如金融工程、计量经济学等, 被用来处理复杂的非正态分布和高维数据集。

- **统计思想**：替换原理，称为矩方法。
 - ▶ 用样本矩替换总体矩，例如样本均值 \bar{X} 估计总体均值 μ
 - ▶ 用样本矩的函数替换相应的总体矩的函数，例如 $B_2 = A_2 - A_1^2 = S_n^2$ 估计总体方差 $\sigma^2 = EX^2 - (EX)^2 = \mu_2 - \mu^2$
 - ▶ 矩法得到的估计称为：矩估计
- **理论依据**：辛钦大数定律，其实质用到格里纹科定理，即用经验分布函数替换总体分布
- **矩估计的优点**：简单、使用场合广泛（总体分布形式未知时也可以使用）

矩估计

例：对某型号的 20 辆汽车记录其每加仑汽油的行驶里程 (km)，观测数据如下：

29.8 27.6 28.3 27.9 30.1 28.7 29.9 28.0
27.9 28.7 28.4 27.2 29.5 28.5 28.0 30.0
29.1 29.8 29.6 26.9,

经计算有

$$\bar{x} = 28.695, \quad s^2 = 0.9185, \quad m_{0.5} = 28.6$$

由此给出总体均值、方差和中位数的估计分别为：28.695, 0.9185 和 28.6。

概率函数已知时未知参数的矩估计

设总体具有已知的概率函数 $P(x; \theta_1, \dots, \theta_k)$, X_1, X_2, \dots, X_n 是样本, 假定总体的 k 阶原点矩 μ_k 存在, 若 $\theta_1, \dots, \theta_k$ 能够表示成 μ_1, \dots, μ_k 的函数 $\theta_j = \theta_j(\mu_1, \dots, \mu_k)$, 则可给出 θ_j 的矩估计为

$$\hat{\theta}_j = \theta_j(a_1, \dots, a_k), \quad j = 1, \dots, k,$$

其中

$$a_j = \frac{1}{n} \sum_{i=1}^n X_i^j.$$

是前 k 阶样本原点矩。

例：设总体服从指数分布，其密度函数为

$$p(x; \lambda) = \lambda e^{-\lambda x}, \quad x > 0,$$

- X_1, \dots, X_n 是样本，由于只有一个未知参数 λ ，故 $k = 1$ 。
- 对于总体，由于 $E(X) = 1/\lambda$ ，即 $\lambda = 1/E(X)$ ，故 λ 的矩估计为

$$\hat{\lambda} = 1/\bar{X}.$$

矩估计的求法

- 另外, 由于 $\text{Var}(X) = 1/\lambda^2$, 其反函数为 $\lambda = 1/\sqrt{\text{Var}(X)}$ 。因此, 从替换原理来看, λ 的矩法估计也可取为

$$\hat{\lambda}_1 = 1/S,$$

其中 S 为样本标准差。

- 矩估计可能是不唯一的, 这是矩法估计的一个缺点。
- 此时通常应该尽量采用低阶矩给出未知参数的估计。

矩估计的求法

例: X_1, X_2, \dots, X_n 是来自 (a, b) 上的均匀分布 $U(a, b)$ 的样本, a 与 b 均是未知参数, 这里 $k = 2$, 由于

$$E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12},$$

不难推出

$$a = E(X) - \sqrt{3 \text{Var}(X)}, \quad b = E(X) + \sqrt{3 \text{Var}(X)}.$$

由此即可得到 a, b 的矩估计:

$$\hat{a} = \bar{X} - \sqrt{3}S, \quad \hat{b} = \bar{X} + \sqrt{3}S$$

概率函数 $P(x, \theta)$ 未知时, 未知参数的矩估计

例: 设总体 X 的均值 μ 和方差 σ^2 存在且未知。 X_1, \dots, X_n 是取自总体 X 的样本。求 μ 和 σ^2 的矩估计。

先求总体矩:

$$\mu_1 = E(X) = \mu,$$

$$\mu_2 = E(X^2) = \text{Var}(X) + E^2(X) = \sigma^2 + \mu^2.$$

再求样本矩:

$$\hat{\mu}_1 = A_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

$$\hat{\mu}_2 = A_2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

概率函数 $P(x, \theta)$ 未知时, 未知参数的矩估计

所以:

$$\begin{cases} \hat{\mu} = \bar{X}, \\ \hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S_n^2. \end{cases}$$

矩估计的性质

- 回顾定理：若总体 X 的 k 阶矩 $E(X^k) = \mu_k$ 存在，则由辛钦大数定律可得

$$A_k \xrightarrow{p} \mu_k$$

进而 $g(A_1, A_2, \dots, A_k) \xrightarrow{p} g(\mu_1, \mu_2, \dots, \mu_k)$, $g(\cdot)$ 为连续函数。

- $\theta = g(\mu_1, \mu_2, \dots, \mu_k)$ 的矩估计 $\hat{\theta} = g(A_1, A_2, \dots, A_k)$ 是相合的。
- 即矩估计一般都具有相合性。

小结

- 1 理解并掌握矩估计的统计思想、理论基础、一般求法及其统计性质。
- 2 了解矩估计的优缺点。

教材习题 6.2: 第 2、3、4、5 题

Outline

- 1 参数估计做什么？
- 2 点估计与估计量的评选标准
- 3 矩估计及其统计性质
- 4 极大似然估计

极大似然法的基本思想

- 高斯（1821）和费希尔（1922）
- 极大似然估计通过观察样本的结果出现的可能（似然）推断总体
- 需要总体的分布已知
- 一般需要样本量较大，因此不讨论无偏性
- 下面通过两个例子叙述最大似然估计的基本思想

极大似然估计举例

例：设有外形完全相同的两个箱子，甲箱中有 99 个白球和 1 个黑球，乙箱中有 99 个黑球和 1 个白球，今随机地抽取一箱，并从中随机抽取一球，结果取得白球。问这球是从哪一个箱子中取出？

由于甲箱中抽出白球的概率为 **0.99**，乙箱中抽出白球的概率为 **0.01**，因此最像（最大似然）从甲箱抽出。

极大似然法的基本思想 (极大似然原理): 若一试验有 n 个可能结果 A_1, \dots, A_n , 现做一试验, 若事件 A_i 发生了, 则认为事件 A_i 在这 n 个可能结果中出现的概率最大。即该试验的条件有利于事件 A_i 的发生。

极大似然估计

- 极大似然估计 (Maximum Likelihood Estimator, MLE) 就是在一次抽样中, 若得到观测值 x_1, x_2, \dots, x_n , 则选取 $\hat{\theta}(x_1, \dots, x_n)$ 作为 θ 的估计值使得当 $\theta = \hat{\theta}(x_1, \dots, x_n)$ 时, 该样本观测值出现的概率最大。
- 当从总体中随机抽取 n 个样本观测值, 最合理的参数估计量应该使得从总体中抽取该 n 个样本观测值的概率最大 (有利于该 n 个样本观测值的出现)。

极大似然估计：举例

- 设产品分为合格和不合格品两类。用随机变量 X 表示某个产品经检查的不合格数，则 $X = 0$ 表示合格品， $X = 1$ 表示不合格品，则 X 服从两点分布 $b(1, p)$ ，其中 p 未知。
- 现抽取 n 个产品看是否合格，得到样本 x_1, \dots, x_n ，这批观测值发生的概率为

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n; p) &= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

- 由于 p 是未知的，根据最大似然原理，我们应选择 p 使得上式表示的概率尽可能大。

- 由于 x_1, \dots, x_n 可观察，可将上式看作是未知参数 p 的函数，用 $L(p)$ 表示，称作**似然函数** (likelihood function):

$$L(p) = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}.$$

- 接下来的任务是找出 p ，使得上式最大。一般做法是将似然函数取对数，并令其一阶导等于 0。

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0$$

解得 p 的最大似然估计为:

$$\hat{p} = \hat{p}(x_1, \dots, x_n) = \sum_{i=1}^n x_i / n = \bar{x}$$

极大似然估计

- 对离散总体，我们可以写出样本观测值出现的概率，此概率依赖于未知参数 θ ，将概率看作是未知参数 θ 的函数，即为似然函数 $L(\theta)$ 。
- 对连续总体，样本观测值出现的概率均为 0，因此需要用联合密度函数表示其似然函数。

极大似然估计

- 设总体的概率函数为 $p(x; \theta)$, $\theta \in \Theta$ 是参数 θ 可能取值的参数空间, x_1, x_2, \dots, x_n 是样本, 将样本的联合概率函数看成 θ 的函数, 用 $L(\theta; x_1, x_2, \dots, x_n)$ 表示, 简记为 $L(\theta)$,

$$L(\theta) = L(\theta; x_1, \dots, x_n) = p(x_1; \theta) \cdots p(x_n; \theta)$$

$L(\theta)$ 称为样本的似然函数。

- 如果某统计量 $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ 满足

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

则称 $\hat{\theta}$ 是 θ 的极大似然估计, 简记为 MLE (Maximum Likelihood Estimate)。

极大似然估计：举例

设一个试验有三种可能的结果，发生的概率分别为

$$p_1 = \theta^2, \quad p_2 = 2\theta(1 - \theta), \quad p_3 = (1 - \theta)^2$$

现做 n 次试验，观测到三种结果发生的次数分别为 n_1, n_2, n_3 ($n_1 + n_2 + n_3 = n$)。

1 似然函数：

$$\begin{aligned} L(\theta) &= (\theta^2)^{n_1} [2\theta(1 - \theta)]^{n_2} [(1 - \theta)^2]^{n_3} \\ &= 2^{n_2} \theta^{2n_1 + n_2} (1 - \theta)^{2n_3 + n_2}. \end{aligned}$$

2 其对数似然函数为

$$\begin{aligned} \ln L(\theta) &= (2n_1 + n_2) \ln \theta + \\ &\quad (2n_3 + n_2) \ln(1 - \theta) + n_2 \ln 2. \end{aligned}$$

3 将之关于 θ 求导，并令其为 0 得到似然方程

$$\frac{2n_1 + n_2}{\theta} - \frac{2n_3 + n_2}{1 - \theta} = 0$$

解得

$$\hat{\theta} = \frac{2n_1 + n_2}{2(n_1 + n_2 + n_3)} = \frac{2n_1 + n_2}{2n}.$$

4 由于

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{2n_1 + n_2}{\theta^2} - \frac{2n_3 + n_2}{(1 - \theta)^2} < 0.$$

所以 $\hat{\theta}$ 是极大值点。

极大似然估计：举例

对正态总体 $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$ 是二维参数, 设有样本 x_1, x_2, \dots, x_n , 则似然函数及其对数分别为

$$\begin{aligned}L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}. \\ \ln L(\mu, \sigma^2) &= -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln(2\pi).\end{aligned}$$

将 $\ln L(\mu, \sigma^2)$ 分别关于两个分量求偏导并令其为 0，即得到似然方程组：

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0.$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0.$$

极大似然估计的一般步骤

1 写出似然函数

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i; \theta_1, \dots, \theta_m).$$

2 写出对数似然函数

$$l(\theta) = \ln L(\theta; x_1, x_2, \dots, x_n) = \sum_{i=1}^n \ln p(x_i; \theta_1, \dots, \theta_m).$$

3 $l(\theta)$ 分别对 θ_j 求偏导, 并令其等于 0, 解得 $\hat{\theta}_1, \dots, \hat{\theta}_m$

$$\frac{\partial l(\theta)}{\partial \theta_j} = 0, j = 1, 2, \dots, m.$$

4 若二阶导数矩阵的非正定, 即可得 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 分别为 $\theta_1, \dots, \theta_m$ 的极大似然估计。

指数分布参数的 MLE

请参考[这里](#).

极大似然估计

虽然求导函数是求极大似然估计最常用的方法，但并不是在所有场合求导都是有效的。

- 设 x_1, x_2, \dots, x_n 是来自均匀总体 $U(0, \theta)$ 的样本，试求 θ 的极大似然估计。
- 似然函数

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^n I_{\{0 \leq x_i \leq \theta\}} = \frac{1}{\theta^n} I_{\{x_{(n)} \leq \theta\}}$$

$x_{(n)}$ 是最大次序统计量。

- 要使 $L(\theta)$ 达到最大，首先一点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 的极大似然估计： $\hat{\theta} = x_{(n)}$ 。

极大似然估计的不变性

- 极大似然估计有一个简单而有用的性质：如果 $\hat{\theta}$ 是 θ 的极大似然估计，则对任一函数 $g(\theta)$ ，其极大似然估计为 $g(\hat{\theta})$ 。该性质称为极大似然估计的不变性。
- 这使一些复杂结构的参数的极大似然估计的获得变得容易了。

极大似然估计的不变性

- 设 X_1, X_2, \dots, X_n 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, μ, σ^2 的极大似然估计为

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

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

于是由不变性可得如下参数的极大似然估计:

- ▶ 总体标准差 σ 的 MLE 是 $\hat{\sigma} = S_n$.
- ▶ 概率 $P(X < 3) = \Phi\left(\frac{3-\mu}{\sigma}\right)$ 的 MLE 是 $\Phi\left(\frac{3-\bar{X}}{S_n}\right)$.
- ▶ 总体 0.90 分位 $x_{0.90} = \mu + \sigma u_{0.90}$ 的 MLE 是 $\bar{X} + S_n \cdot u_{0.90}$, 其中 $u_{0.90}$ 为标准正态分布的 0.90 分位数。

极大似然估计的渐近正态性

设总体 X 有密度函数 $p(X; \theta)$, 若

1 对任意的 x , $\frac{\partial \ln p}{\partial \theta}$, $\frac{\partial^2 \ln p}{\partial \theta^2}$, $\frac{\partial^3 \ln p}{\partial \theta^3}$ 对所有 θ 都存在;

2 对任意 θ , 有

$$\left| \frac{\partial p}{\partial \theta} \right| < F_1(x), \quad \left| \frac{\partial^2 p}{\partial \theta^2} \right| < F_2(x), \quad \left| \frac{\partial^3 \ln p}{\partial \theta^3} \right| < F_3(x),$$

其中函数 $F_1(x)$, $F_2(x)$, $F_3(x)$ 满足

$$\int_{-\infty}^{\infty} F_1(x) dx < \infty, \quad \int_{-\infty}^{\infty} F_2(x) dx < \infty, \quad \sup_{\theta \in \Theta} \int_{-\infty}^{\infty} F_3(x) p(x; \theta) dx < \infty;$$

3 对任意 θ , $0 < \mathcal{I}(\theta) \equiv \int_{-\infty}^{\infty} \left(\frac{\partial \ln p}{\partial \theta} \right)^2 p(x; \theta) dx < \infty$.

极大似然估计的渐近正态性

那么 θ 的极大似然估计具有相合性和渐近正态性

$$\hat{\theta}_n \sim AN\left(\theta, \frac{1}{n\mathcal{I}(\theta)}\right)$$

其中

$$\mathcal{I}(\theta) \equiv E\left[\frac{\partial}{\partial\theta} \ln p(X; \theta)\right]^2 = \int_{-\infty}^{\infty} \left(\frac{\partial \ln p}{\partial\theta}\right)^2 p(x; \theta) dx$$

称为**费希尔信息量 (Fisher Information)**.

Fisher 信息量

- 定义得分: $\frac{\partial}{\partial \theta} \log p(\mathbf{X}; \theta)$.
- 得分的期望:

$$\begin{aligned} \mathbb{E} \left[\frac{\partial}{\partial \theta} \ln p(\mathbf{X}; \theta) \right] &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial \theta} p(\mathbf{x}; \theta)}{p(\mathbf{x}; \theta)} p(\mathbf{x}; \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} p(\mathbf{x}; \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \mathbf{1} \\ &= \mathbf{0}. \end{aligned}$$

- 得分的方差:

$$\mathcal{I}(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log p(\mathbf{X}; \theta) \right)^2 \right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta} \log p(\mathbf{x}; \theta) \right)^2 p(\mathbf{x}; \theta) d\mathbf{x}$$

- $\mathcal{I}(\theta)$ 越大, 说明得分的绝对值越高, 总体分布中包含未知参数 θ 的信息越多 (进而 MLE 的方差越小)。
- $\mathcal{I}(\theta)$ 还可以表示为 $\mathcal{I}(\theta) = -\mathbb{E} \frac{\partial^2 \ln p(\mathbf{X}; \theta)}{\partial \theta^2}$ 。

MLE 的渐近分布：举例

- 设 x_1, \dots, x_n 是来自 $N(\mu, \sigma^2)$ 的样本，可以验证该总体分布在 σ^2 已知或 μ 已知时，满足上述三个正则条件。
- 在 σ^2 已知时， μ 的 MLE 为 $\hat{\mu} = \bar{x}$ ， $\hat{\mu}$ 服从渐近正态分布，下面求 $\mathcal{I}(\mu)$ 。

$$\ln p(\mathbf{x}) = -\ln \sqrt{2\pi} - \frac{1}{2} \ln \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}$$

$$\frac{\partial \ln p}{\partial \mu} = \frac{x - \mu}{\sigma^2}$$

$$\mathcal{I}(\mu) = E \left(\frac{x - \mu}{\sigma^2} \right)^2 = \frac{1}{\sigma^2}$$

从而有

$$\hat{\mu} \sim AN(\mu, \sigma^2/n).$$

MLE 的渐近分布：举例

- 在 μ 已知时, σ^2 的 MLE 为 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$, 下面求 $\mathcal{I}(\sigma^2)$ 。

$$\begin{aligned}\frac{\partial \ln p}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x - \mu)^2 = \frac{(x - \mu)^2 - \sigma^2}{2\sigma^4} \\ \mathcal{I}(\sigma^2) &= \frac{E[(x - \mu)^2 - \sigma^2]^2}{4\sigma^8} \\ &= \frac{\text{Var}((x - \mu)^2)}{4\sigma^8} = \frac{1}{2\sigma^4}\end{aligned}$$

从而

$$\hat{\sigma}^2 \sim AN(\sigma^2, 2\sigma^4/n).$$

随堂练习三：求费希尔信息量

- 1 设总体分布为指数分布，其密度函数为

$$p(x; \lambda) = \lambda e^{-\lambda x}, x > 0, \lambda > 0,$$

试求总体分布的费希尔信息量。

- 2 设总体分布为泊松分布 $P(\lambda)$ ，其分布列为

$$p(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots,$$

试求总体分布的费希尔信息量。

- 3 验证 $\mathcal{I}(\theta) = -E \frac{\partial^2 \ln p(x; \theta)}{\partial \theta^2} = E \left(\frac{\partial \ln p(x; \theta)}{\partial \theta} \right)^2$ 。

- 理解并熟悉掌握极大似然原理, 似然函数、对数似然函数、似然方程(组) 含义, 及其极大似然估计的优缺点等。
- 熟悉掌握极大似然估计一般步骤。
- 理解并熟悉极大似然估计的统计性质 (不变性、相合性、渐近正态性、有效性)。
- 掌握如何计算不同分布下的极大似然估计、Fisher 信息量等。

教材习题 6.3: 第 1、2、4、7、8 题