

数学基础班第 4 课课件：参数估计

管枫

七月在线

July, 2016

主要内容

- 点估计
 - 矩估计
 - 极大似然估计
 - 点估计的评判准则
- 区间估计
 - 置信区间

本次课程参考中科大统计系张卫平老师的课程材料:

[http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/,lec\(14,15\)](http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/,lec(14,15))

[http://staff.ustc.edu.cn/~zwp/teach/Prob-Stat/,lec\(4,5,6,7,8\)](http://staff.ustc.edu.cn/~zwp/teach/Prob-Stat/,lec(4,5,6,7,8))

- 本节课常用数学记号

$\mathbb{R}^n$ 实坐标空间

$f(x)$ 函数

$\int_a^b f(x)dx$ 函数积分

X 随机变量

$E(X)$ 随机变量的期望

$Var(x)$ 随机变量的方差

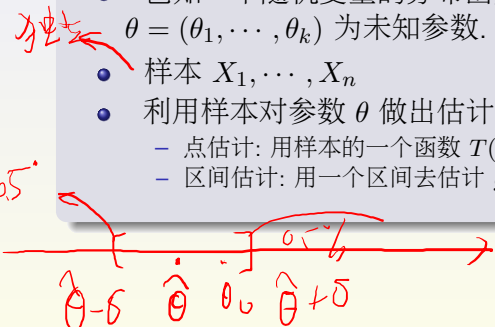
$\alpha_n(x)$ 随机变量的 n 阶原点矩

$\mu_n(x)$ 随机变量的 n 阶中心矩

参数估计

参数估计问题

- 已知一个随机变量的分布函数 $X \sim f_{\theta}(x)$, 其中 $\theta = (\theta_1, \dots, \theta_k)$ 为未知参数.
- 样本 $X_1, \dots, X_n$
- 利用样本对参数 θ 做出估计, 或者估计 θ 的某个函数 $g(\theta)$
 - 点估计: 用样本的一个函数 $T(X_1, \dots, X_n)$ 去估计 $g(\theta)$
 - 区间估计: 用一个区间去估计 $g(\theta)$



点估计：矩估计

矩估计

- 矩估计法的基本思想是根据大数定律，利用样本矩对总体分布矩进行估计.
- 然后利用总体矩与参数的关系来对参数进行估计.

记号:

样本 k 阶矩:
$$a_k(X) = \frac{1}{n} \sum_{i=1}^n X_i^k \quad m_k(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

总体 k 阶矩:
$$\alpha_k(X) = E(X^k) \quad \mu_k(X) = E((X - E(X))^k)$$

点估计：矩估计

矩估计的基本原理：大数定律

根据大数定律我们知道, 对于任何随机变量 X , 当样本数 $n \rightarrow \infty$ 时, $\frac{1}{n} \sum_{i=1}^n X_i$ 收敛于 $E(X)$. 所以

$$a_1(X) \rightarrow \alpha_1(X)$$

对于任意的 k 阶矩, 令 $Y = X^k$, 那么 Y 也是一个随机变量, 所以同样满足大数定律, 于是

$$a_k(X) = a_1(Y) \rightarrow \alpha_1(Y) = \alpha_k(X)$$

而中心矩都可以表示成原点矩的多项式, 所以我们同样有

$$m_k(X) \rightarrow \mu_k(X)$$

点估计：矩估计

Example (两点分布的参数估计)

X 服从两点分布取值为 $\{-1, 1\}$, $P(-1) = 1 - \theta$, $P(1) = \theta$. 现在独立重复实验 n 次, 得到样本 $X_1, \dots, X_n$. 请利用矩估计来估计参数 θ .

首先考虑哪一个矩可以用来估计参数 θ . 对于两点分布来说

$$E(X) = (1 - \theta) \cdot (-1) + \theta \cdot 1 = 2\theta - 1$$

$$E(X^2) = (1 - \theta) \cdot 1 + \theta \cdot 1 = 1$$

我们看到一阶矩 $E(X)$ 与 θ 有简单直接的关系 $\theta = \frac{1+E(X)}{2}$
所以我们使用一阶样本矩估计. 得到一个参数估计量 $\hat{\theta} = \frac{1+\bar{X}}{2}$.

点估计：矩估计

Example (正态分布的参数估计)

X 服从参数为 $\theta = (\mu, \sigma)$ 的正态分布，独立重复实验 n 次得到样本 $X_1, \dots, X_n$. 请利用矩估计来估计参数 θ .

首先考虑哪一个矩可以用来估计参数 θ . 对于正态分布来说

$$\begin{aligned}E(X) &= \mu \\E(X^2) &= \mu^2 + \sigma^2 \\E(X^3) &= \frac{1}{6}(3\mu\sigma^2 + \mu^3)\end{aligned}$$

原则上 2 阶矩和 3 阶矩都可以用来估计 σ^2 ，所以矩估计一般来讲是不唯一的。当有多种选择的时候，我们尽可能选择阶数较小的矩。所以可以采用

$$\hat{\mu} = a_1(X), \hat{\sigma}^2 = m_2(X).$$

此处对 σ 的估计并不是最优的，在点估计的评判标准那一节，我们进行更细致的讨论。

点估计：极大似然估计

极大似然估计

- 给定随机变量的分布与未知参数，利用观测到的样本计算似然函数
- 选择最大化似然函数的参数作为参数估计量.

点估计：极大似然估计

独立

极大似然估计基本原理：最大化似然函数

假设样本 $\{X_1, \dots, X_n\}$ 服从概率密度函数 $f_\theta(x)$. 其中 $\theta = (\theta_1, \dots, \theta_k)$ 是未知参数.

当固定 x 的时候, $f_\theta(x)$ 就是 θ 的函数, 我们把这个函数称为似然函数, 记为 $L_x(\theta)$ 或 $L(\theta)$.

似然函数不是概率, 但是很类似于概率. 当 θ 给定的时候, 它是概率密度. 当 x 给定, θ 变化的时候, 他就类似于在表示, 在这个观测量 x 的条件下, 参数等于 θ 的可能性 (不是概率). 起个名字叫做似然函数.

$$L_x(\theta_1) \neq L_x(\theta_2)$$

$$P(\theta = \theta_1) \neq P(\theta = \theta_2)$$

点估计：极大似然估计

极大似然估计基本原理：最大化似然函数

假设 $x = (x_1, \dots, x_n)$ 是样本的观测值. 那么整个样本的似然函数就是

$$f_{\theta}(x) = f_{\theta}(x_1) \cdots f_{\theta}(x_n) = L_{x_1}(\theta) \cdots L_{x_n}(\theta)$$
$$\underline{L_x(\theta)} = \prod_{i=1}^n \underline{L_{x_i}(\theta)}$$

这是一个关于 θ 的函数, 选取使得 $L_x(\theta)$ 最大化的 $\hat{\theta}$ 作为 θ 的估计量.

最大化似然函数 θ , 相当于最大化似然函数的对数

$l_x(\theta) = \ln(L_x(\theta))$. 一般我们求解似然函数或者对数似然函数的驻点方程

$$\frac{\partial l(\theta)}{\partial \theta} = 0, \quad \theta \in \{1, \dots, k\}$$
$$\frac{dl(\theta)}{d\theta} = 0, \quad (\text{或者 } \frac{dL(\theta)}{d\theta} = 0)$$

然后判断整个驻点是否最大点.(求驻点可以用牛顿法, 或者梯度法等等)

点估计：极大似然估计

Example (正态分布的参数估计)

X 服从参数为 $\theta = (\mu, \sigma)$ 的正态分布，独立重复实验 n 次得到样本 $X_1, \dots, X_n$. 请利用极大似然估计来估计参数 θ .

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

$$l(\mu, \sigma) = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2} \ln(\sigma^2)$$

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

$$\begin{aligned}\sum_{i=1}^n x_i &= n\mu \\ \mu &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \bar{x}\end{aligned}$$

所以似然方程为 $\frac{\partial l}{\partial \sigma} = \frac{\partial l}{\partial \mu} = 0$, 也就是

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

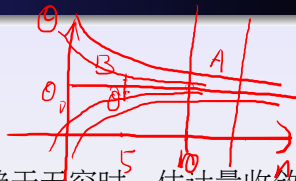
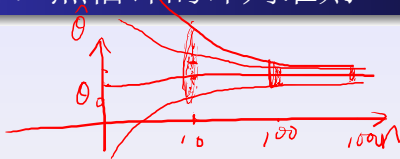
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

因此得到极大似然估计量 $\sum (x_1, \dots, x_n) / n$

$$\hat{\mu} = \bar{X}$$

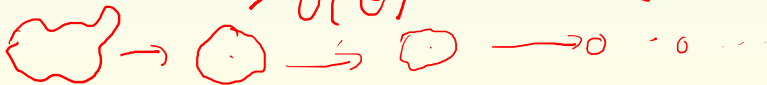
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

点估计: 点估计的评判准则



- 相合性 (consistency): 当样本数量趋于无穷时, 估计量收敛于参数真实值.
- 无偏性 (bias): 对于有限的样本, 估计量所符合的分布之期望等于参数真实值.
- 有效性 (efficiency): 估计值所满足的分布方差越小越好.
- 渐进正态性 (asymptotic normality): 当样本趋于无穷时, 去中心化去量纲化的估计量符合标准正态分布.

$$\frac{[\hat{\theta} - \mu(\theta)]}{\sigma(\hat{\theta})} \rightarrow N(0, 1)$$



点估计：相合性

相合性

相合性是最基本的要求，矩估计的相合型是由大数定律来保证的。极大似然估计的相合型也是隐含的由大数定律来保证的

假设一个随机变量 X 服从 $f_{\theta_0}(x)$ 。最大化 $l_x(\theta)$ 跟最大化 $\frac{1}{n}l_x(\theta)$ 是一样的

$$X = (X_1, \dots, X_n)$$

$$\frac{1}{n}l_x(\theta) = \frac{1}{n} \sum_{i=1}^n l_{x_i}(\theta)$$

$$= \frac{1}{n} \sum_{i=1}^n \ln(f_{\theta}(x_i))$$

$$L_X(\theta) = \prod_{i=1}^n L_{x_i}(\theta)$$
$$l_X(\theta) = \sum_{i=1}^n l_{x_i}(\theta)$$

这个无穷求和就收敛于（大数定律）

$$\underline{E(\ln(f_{\theta}(X)))} = \int \underline{\ln(f_{\theta}(x))} \underline{f_{\theta_0}(x)dx}$$

点估计：相合性

而 $\hat{\theta}$ 是 $\frac{1}{n}l_x(\theta)$ 的极大值点，所以 $\lim \hat{\theta}$ 收敛于 $E(\ln(f_\theta(X)))$ 的极大值点。

所以我们只需要证明 θ_0 确实是 $E(\ln(f_\theta(X)))$ 的极大值点。因为 $\ln(x)$ 是个凹函数，根据琴生不等式我们有

$$\begin{aligned} & \int_x \ln(f_\theta(x)) f_{\theta_0}(x) dx - \int_x \ln(f_{\theta_0}(x)) f_{\theta_0}(x) dx \\ &= \int_x \ln(f_\theta(x)/f_{\theta_0}(x)) f_{\theta_0}(x) dx \\ &\leq \ln\left(\int_x \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x) dx\right) \\ &= \ln\left(\int_x f_\theta(x) dx\right) \\ &= \ln(1) = 0 \end{aligned}$$

$$E(\ln(f_{\theta_0}(X))) \geq E(\ln(f_\theta(X)))$$

$$\ln A - \ln B = \ln \frac{A}{B}$$



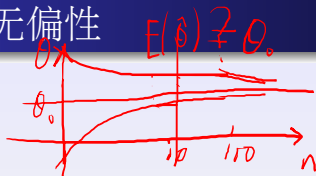
点估计：相合性

也就是说

$$E(\ln(f_{\theta}(X))) - E(\ln(f_{\theta_0}(X))) \leq 0$$

于是 θ_0 就是关于 θ 的函数 $E(\ln(f_{\theta}(X)))$ 的极大值点.

点估计：无偏性



无偏性

任何一个满足相合性的参数估计，当样本趋于无穷的时候都会收敛与参数的真实值. 但是对于有限样本的情况下，这个估计值的期望不见得总等于参数的真实值.

比如我们做的关于正态分布的方差的估计 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$.

我们计算一下等式右边的期望值

点估计：无偏性

$$E \frac{\sum_{i=1}^n (x_i - \mu + \mu - \bar{X})^2}{n} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} + \frac{\sum_{i=1}^n (\mu - \bar{X})^2}{n} + \frac{\sum_{i=1}^n 2(x_i - \mu)(\mu - \bar{X})}{n}$$

$$= E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right) + E(\mu - \bar{X})^2 +$$

$$E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2\right)$$

$$\begin{aligned} & E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2\right) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 + (\mu - \bar{X})^2 + 2(\mu - \bar{X}) \sum_{i=1}^n (X_i - \mu)\right) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) + E((\mu - \bar{X})^2) + 2E((\mu - \bar{X}) \sum_{i=1}^n (X_i - \mu)) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) + E((\mu - \bar{X})^2) + 2E((\mu - \bar{X}) \sum_{i=1}^n (X_i - \mu)) \\ &= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) + E((\mu - \bar{X})^2) + 2E((\mu - \bar{X}) \sum_{i=1}^n (X_i - \mu)) \end{aligned}$$

$$= E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right) + E((\mu - \bar{X})^2)$$

$$= E((X_i - \mu)^2) - E((\mu - \bar{X})^2)$$

$$= \sigma^2 - \text{Var}(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

所以我们倾向于“低估” σ^2 . 那么我们低估的这个值 $\text{Var}(\bar{X})$ 等于多少呢?

点估计：无偏性

令 $\underline{Y_i = X_i - \mu}$, 那么 $\bar{X} - \mu = \bar{Y}$, 所以

$$E((\mu - \bar{X})^2) = E((\bar{Y})^2)$$

Y 的特征函数是 $\phi_Y(t) = e^{-\frac{t^2 \sigma^2}{2}}$, 所以

$$\begin{aligned}\phi_{\bar{Y}}(t) &= (e^{-\frac{t^2 \sigma^2}{2n^2}})^n \\ &= e^{-\frac{t^2 (\sigma^2 / \sqrt{n})^2}{2}}\end{aligned}$$

于是 $Var(\bar{X}) = \underline{Var(\bar{Y})} = \underline{\sigma^2/n}$. 所以

$$E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{n-1}{n} \sigma^2$$

因此 $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 才是 σ^2 的无偏估计量.

$$\begin{aligned}\phi_{\bar{Y}}(t) &= E(e^{it\bar{Y}}) \\ &= E\left(e^{i\frac{t}{n} \sum_{k=1}^n Y_k}\right) \\ &= E\left(\prod_{k=1}^n e^{i\frac{t}{n} Y_k}\right) \\ &= \prod_{k=1}^n E\left(e^{i\frac{t}{n} Y_k}\right) \\ &= \prod_{k=1}^n \phi_{Y_k}\left(\frac{t}{n}\right) \\ &= \left(\phi_Y\left(\frac{t}{n}\right)\right)^n\end{aligned}$$

点估计：有效性

$$(1^2 + \dots + 1^2)(w_1^2 + \dots + w_n^2) \geq (w_1 + \dots + w_n)^2$$
$$n \sum_{i=1}^n w_i^2 \geq 1 \Rightarrow \sum_{i=1}^n w_i^2 \geq \frac{1}{n}$$

有效性

如果两个参数估计量 $\hat{\theta}$ 和 $\tilde{\theta}$ ，既是相合的又是无偏的，那么他们两个中方差较小的那一个比较好。如果

$$\text{Var}(\hat{\theta}) \geq \text{Var}(\tilde{\theta})$$

那么我们就认为 $\tilde{\theta}$ 比较有效。

此处举一个例子。

$$\frac{\sigma^2}{n} \leq \sum_{i=1}^n \sigma^2 w_i^2$$
$$\frac{1}{n} \leq \sum_{i=1}^n w_i^2$$
$$(a_1^2 + \dots + a_n^2)(b_1^2 + \dots + b_n^2) \geq (a_1 b_1 + \dots + a_n b_n)^2$$
$$\text{令 } a_i = 1, \quad b_i = w_i$$

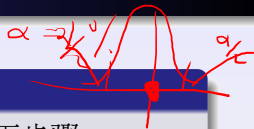
点估计：渐进正态性

渐进正态性

渐进正态这里涉及到一些矩阵的知识, 尤其是极大似然估计的渐进正态性质. 等讲完线性代数, 在作为一个例子扩展讲解一下。

区间估计：置信区间

$$\bar{x} = 2\theta - 1 \quad \theta = \frac{1 + \bar{x}}{2}$$



置信区间

置信区间可以认为是点估计的一个扩展。分为如下步骤

- 找到一个点估计 T
- 找出一个 T 与 θ 的函数满足某一个已知的分布 (F)
- 利用这个已知的分布 F 的 $\alpha/2$ 分位数，来求出参数的置信区间。

如果这个分布 F 很难找到，那么还有一种近似的方法

- 找到一个点估计 T
- 利用渐进正态的性质，发现 T 在 n 很大的时候满足某种正态分布。
- 利用这个已知的正态分布的 $\alpha/2$ 分位数，来求出参数的置信区间。

$$\theta_{\sqrt{n}} \rightarrow N(0, 1)$$

$$\{f_\theta(x) \quad x_1, \dots, x_n$$

$$S = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\theta} \sim N(\mu, \sigma_{\hat{\theta}})$$

$$\alpha = 1\%, \quad 99\%$$

$$[\hat{\theta} - k \frac{\sigma}{\sqrt{n}}, \hat{\theta} + k \frac{\sigma}{\sqrt{n}}], \quad \hat{\theta}, \quad \sigma(\hat{\theta}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S}{\sqrt{n}}$$

谢谢大家!

$$\begin{matrix} x_1 \rightarrow x_1 \\ \vdots \\ x_n \rightarrow x_n \end{matrix} \rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i \rightarrow \text{点估计}$$

$$[\hat{\theta} - k \frac{S}{\sqrt{n}}, \hat{\theta} + k \frac{S}{\sqrt{n}}] = I$$

$$\underbrace{99\%}_{\theta_0} \in I$$