统计基本目的：从样本出发推断总体分布. ——统计推断(statistical inference)

$$\text{Statistical Inference}\begin{cases}\text{Sampling distribution}\\ \text{Estimation}\begin{cases}\text{Point Estimation}\\ \text{Interval Estimation}\end{cases}\\ \text{Hypothesis Testing}\end{cases}$$

第三章　点估计

§3.1 引言

一、参数

参数估计中的"参数"是泛指刻画总体某方面性质、特征等的值.

对参数分布族$\{\mathcal{F}_\theta : \theta \in \Theta\}$, 其中$\theta$未知. 那么$\theta$是参数.

有时,我们不一定想知道θ本身. 例如对于来自正态分布族$\{N(\mu, \sigma^2), -\infty < \mu < +\infty, \sigma > 0\}$的总体$X$而言, 未知参数是$\theta = (\mu, \sigma)$, 常常我们可能只对$\mu$感兴趣,不需要知道$\sigma$; 也可能对$\mu$和$\sigma$都感兴趣, 也有可能对变异系数$\sigma/\mu$, $P(X > 1)$等感兴趣. 在这些统计研究中研究者所感兴趣的值, 称为待估参数, 简称参数.

对于参数模型而言, 待估参数是分布参数的函数$g(\theta)$.

对于非参数分布族$\mathcal{F}=\{F\}$, 我们仍然有刻画总体某些方面性质的量(例如:总体均值,总体方差,中位数等), 需要估计它们的值, 这些量也是待估参数. 如果知道了总体分布函数$F(x)$, 这些参数的值就可以通过总体分布F计算出来, 因此待估参数是总体分布的泛函$g(F)$, 或写为θ_F.

总而言之, 参数估计中的"参数"是泛指刻画总体某方面性质、特征等的值.

为了方便起见, 后续课件中, 有时简写成θ, 此处的θ是广义的参数.

二、参数估计问题

点估计——用一个的具体数值(样本观察值的函数)去估计一个未知参数.

区间估计——-把待估参数估计在某两个界限(都是样本观察值的函数)之间.

三、点估计的概念

Definition

定义3.1.1 用来估计未知参数θ的统计量

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \cdots, X_n)$$

称为参数θ的点估计量(point estimator). 而样本观察值的函数$\widehat{\theta} = \widehat{\theta}(x_1, x_2, \cdots, x_n)$称为参数的估计值.

估计量是一个估计θ的值的规则, 没有样本观察值时也可构造估计量.一旦有了样本观察值, 我们就可以根据这个规则得出θ的估计值.

$$\text{Point Estimation} \begin{cases} \text{Method of Evaluating Estimators} \\ \text{Method of Finding Estimators} \end{cases}$$

四、评价估计量的基本准则

1. 无偏性 (unbiasedness)(反映估计量的平均值)

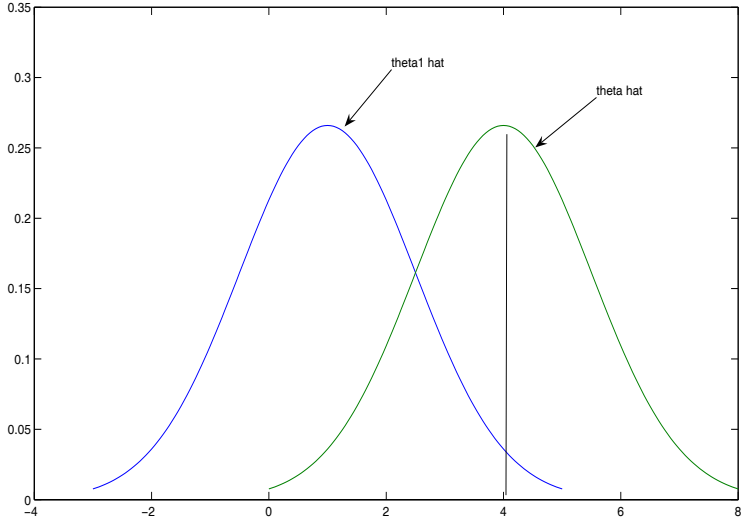
在平均意义下, $\widehat{\theta}$的值不偏离θ.

Definition

定义3.1.2 设$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \cdots, X_n)$是参数$\theta$的一个估计量. 假如

$$\mathsf{E}(\widehat{\theta}) = \theta, \quad \forall \theta \in \Theta,$$

则称$\widehat{\theta}$为θ的无偏估计量(unbiased estimator). 这里Θ是参数空间.


theta1 hat
theta hat
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−2
0
2
4
6
8

E的计算和总体分布F有关, θ也是F的函数. 定义中的表达式应写为

$$\mathsf{E}_F(\widehat{\theta}) = \theta_F, \quad \forall F \in \mathcal{F}$$

或者

$$\mathsf{E}_\theta(\widehat{\theta}) = \theta, \quad \forall \theta \in \Theta.$$

E_θ表示对参数θ所对应的分布F求数学期望.

通常

$$Bias_\theta\widehat{\theta} = \mathsf{E}_\theta(\widehat{\theta}) - \theta$$

称为参数θ的估计$\widehat{\theta}$的偏差.

无偏估计量没有系统偏差.

Example

设$X_1, X_2, \cdots, X_n$是取自均值为μ,方差为σ^2的总体的一个样本. 样本均值为$\overline{X}$, 样本方差为S^2.

由于

$$\mathsf{E}\overline{X} = \frac{\mathsf{E}X_1 + \mathsf{E}X_2 + \cdots + \mathsf{E}X_n}{n} = \mu,$$

$\overline{X}$是总体均值μ的无偏估计. 且

$$\mathsf{E}S^2 = \sigma^2,$$

故S^2是σ^2的无偏估计量.

而考察$S_n^2 = \frac{n-1}{n}S^2$,

$$\begin{aligned}
\mathsf{E}S_n^2 =& \mathsf{E}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
=& \mathsf{E}\frac{1}{n}\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2\right] \\
=& \sigma^2 - \frac{1}{n^2}\mathsf{Var}(X_1+X_2+\cdots+X_n) = \sigma^2 - \frac{\sigma^2}{n} \\
=& \frac{n-1}{n}\sigma^2 \neq \sigma^2.
\end{aligned}$$

S_n^2不是σ^2的无偏估计.

$$Bias_{\sigma^2}S_n^2 = -\frac{1}{n}\sigma^2.$$

纠偏方法:

若$\mathsf{E}(\widehat{\theta}) = a\theta + b, a \neq 0$, a, b为常数, 则

$$\frac{1}{a}(\widehat{\theta} - b)$$

为θ的无偏估计量.

渐近无偏性(asymptotic unbiasedness)

(针对大样本情形)

Definition

设对每个自然数n, $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \cdots, X_n)$是$\theta$的估计量,假如

$$Bias_\theta(\widehat{\theta}_n) = \mathsf{E}(\widehat{\theta}_n) - \theta \to 0, \quad \text{as} \quad n \to \infty, \quad \forall \theta \in \Theta,$$

则称$\widehat{\theta}_n$是θ的渐近无偏估计.

2. 有效性 (efficiency)

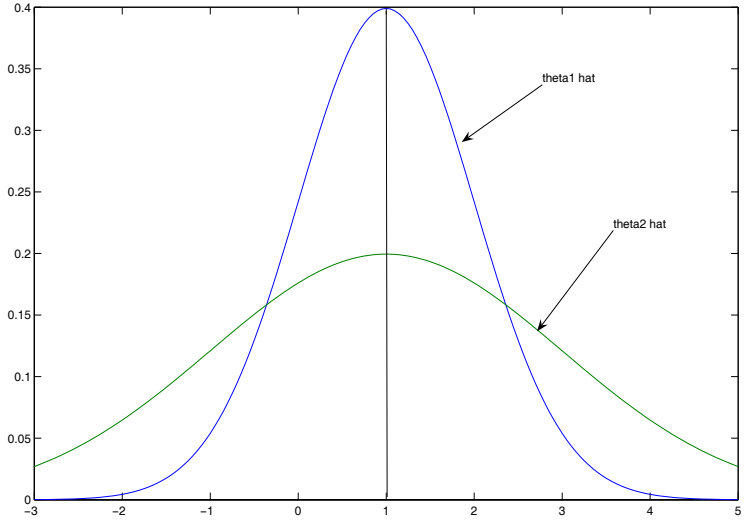
(反映估计量的平均偏离程度)

Definition

定义3.1.3 设$\widehat{\theta}_1 = \widehat{\theta}_1(X_1, X_2, \cdots, X_n)$ 和$\widehat{\theta}_2 = \widehat{\theta}_2(X_1, X_2, \cdots, X_n)$是参数$\theta$的两个无偏估计, 假如

$$\mathsf{Var}_\theta(\widehat{\theta}_1) \leq \mathsf{Var}_\theta(\widehat{\theta}_2), \quad \forall \theta \in \Theta,$$

且至少存在一个$\theta \in \Theta$,使得严格不等号成立,则称$\widehat{\theta}_1$比$\widehat{\theta}_2$有效.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−3
−2
−1
0
1
2
3
4
5
theta1 hat
theta2 hat

3. 均方误差 (Mean Squared Error, MSE)

Definition

定义 设$\widehat{\theta}$是参数θ的一个估计量, 其均方误差定义为

$$\mathrm{MSE}_\theta(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2.$$

自然，我们希望估计的均方误差越小越好.

$$\begin{aligned}\mathrm{MSE}_\theta(\widehat{\theta}) &= \mathsf{E}((\widehat{\theta} - \mathsf{E}\widehat{\theta}) + (\mathsf{E}\widehat{\theta} - \theta))^2 \\ &= \mathsf{Var}(\widehat{\theta}) + (\mathsf{E}\widehat{\theta} - \theta)^2.\end{aligned}$$

$$
\begin{aligned}
&\mathrm{MSE}_\theta(\hat{\theta}) = \mathsf{E}\left[(\hat{\theta}-\theta)^2\right] = \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}+\mathsf{E}\widehat{\theta}-\theta\right)^2\right] \\
&= \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)^2 + 2\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)\left(\mathsf{E}\widehat{\theta}-\theta\right) + \left(\mathsf{E}\widehat{\theta}-\theta\right)^2\right] \\
&= \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)^2\right] + \mathsf{E}\left[2\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)\left(\mathsf{E}\widehat{\theta}-\theta\right)\right] + \mathsf{E}\left[\left(\mathsf{E}\widehat{\theta}-\theta\right)^2\right] \\
&= \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)^2\right] + 2\left(\mathsf{E}\widehat{\theta}-\theta\right)\mathsf{E}\left[\hat{\theta}-\mathsf{E}\widehat{\theta}\right] + \left(\mathsf{E}\widehat{\theta}-\theta\right)^2 \\
&= \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)^2\right] + 2\left(\mathsf{E}\widehat{\theta}-\theta\right)\left(\mathsf{E}\widehat{\theta}-\mathsf{E}\widehat{\theta}\right) + \left(\mathsf{E}\widehat{\theta}-\theta\right)^2 \\
&= \mathsf{E}\left[\left(\hat{\theta}-\mathsf{E}\widehat{\theta}\right)^2\right] + 0 + \left(\mathsf{E}\widehat{\theta}-\theta\right)^2 = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2
\end{aligned}
$$

Example

在正态分布族$\{N(\mu,\sigma^2):-\infty<\mu<+\infty,\sigma>0\}$下, 试用均方误差准则, 对于样本方差$S^2$和样本二阶中心矩$m_{n,2}$来估计$\sigma^2$时的优劣进行评价, 其中样本容量$n\geq 2$.

由于在正态总体下, 有$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$, 又$S^2$为$\sigma^2$的无偏估计, 故

$$\begin{aligned}\mathrm{MSE}_{\sigma^2}(S^2)&=\mathrm{Var}(S^2)=\mathrm{Var}(\frac{(n-1)S^2}{\sigma^2})\cdot\left(\frac{\sigma^2}{n-1}\right)^2\\&=2(n-1)\cdot\frac{\sigma^4}{(n-1)^2}=\frac{2\sigma^4}{n-1}.\end{aligned}$$

样本二阶中心矩$m_{n,2}$的均方误差则为

$$\begin{aligned}
&\mathrm{MSE}_{\sigma^2}(m_{n,2}) = \mathrm{Var}(m_{n,2}) + \mathrm{Bias}(m_{n,2})^2 \\
&= \mathrm{Var}(\frac{n-1}{n}S^2) + (\mathsf{E}(m_{n,2}) - \sigma^2)^2 \\
&= (\frac{n-1}{n})^2 \cdot \frac{2\sigma^4}{n-1} + (\frac{n-1}{n}\sigma^2 - \sigma^2)^2 \\
&= \frac{2n-1}{n^2}\sigma^4.
\end{aligned}$$

由于当$n \geq 2$时, $\frac{2n-1}{n^2} < \frac{2}{n-1}$, 故对于估计$\sigma^2$, 在均方误差准则下, 样本二阶中心矩$m_{n,2}$优于样本方差$S^2$.

4. 相合性(Consistency) (针对大样本情形)

Definition

定义3.1.4 设对每个自然数n, $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \cdots, X_n)$是$\theta$的估计量,假如

$$\widehat{\theta}_n \xrightarrow{P} \theta, \quad n \to \infty, \quad \forall \theta \in \Theta,$$

即对任意的$\epsilon > 0$,

$$\mathsf{P}_\theta(|\widehat{\theta}_n - \theta| \geq \epsilon) \to 0, \quad n \to \infty, \quad \forall \theta \in \Theta,$$

则称$\widehat{\theta}_n$是θ的(弱)相合估计,亦称一致估计.

Definition

若

$$\mathsf{P}_\theta(\lim_{n\to\infty}\widehat{\theta}_n=\theta)=1,\quad \forall\theta\in\Theta,$$

则称$\widehat{\theta}_n$是θ的强相合估计.

若

$$\lim_{n\to\infty}\mathsf{E}_\theta|\widehat{\theta}_n-\theta|^r=0,\quad \forall\theta\in\Theta,$$

则称$\widehat{\theta}_n$是θ的r阶矩相合估计,当$r=2$时,称为均方相合估计.

5. 渐近正态性 (asymptotic normality)

(针对大样本情形)

Definition

设对每个自然数n, $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \cdots, X_n)$是$\theta$的估计量,假如

$$\sqrt{k_n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{D} N(0, \sigma^2), \quad \text{as} \quad n \to \infty, \quad \forall \theta \in \Theta$$

则称$\widehat{\theta}_n$具有渐近正态性, σ^2称为渐近方差.若该估计量同时还是θ的弱相合估计量, 则称$\widehat{\theta}_n$为θ的相合渐近正态估计.

估计方法

点估计方法
- 矩法
- 极大似然法
- 贝叶斯法(Bayes法)
- 最小二乘法 →
  - 最小一乘法
  - M估计
  - S估计　…
- UMVUE
- ⋮

§3.2 估计方法之一——矩法估计 (Method of moments)

一、矩法的统计思想

总体k阶矩

$$\mu_k = \mathsf{E}(X^k) = \int x^k dF(x).$$

将$F(x)$用经验分布函数$F_n(x)$“替换”得样本k阶矩

$$a_{n,k} = \int x^k dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

同样在总体k阶中心矩

$$\nu_k = \mathsf{E}(X - \mathsf{E}X)^k = \int \left(x - \int x dF(x) \right)^k dF(x)$$

中将$F(x)$用经验分布函数$F_n(x)$“替换”得样本k阶中心矩

$$m_{n,k} = \int \left(x - \int x dF_n(x) \right)^k dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k.$$

由于

$$\nu_k = \sum_{r=0}^{k} (-1)^{k-r} \binom{k}{r} \mu_r \mu_1^{k-r},$$

$$\mu_k = \sum_{r=0}^{k} \binom{k}{r} \nu_r \mu_1^{k-r}.$$

中心矩和原点矩之间可以相互转换.

这种用$F_n(x)$替换$F(x)$, 或者说

用样本矩替换相应的总体矩,

即

$$\text{用}a_{n,k}\text{替换}\mu_k, \qquad \text{用}m_{n,k}\text{替换}\nu_k,$$

此时所得估计称为矩法估计.

Definition

设有总体分布族$\{F_\theta;\theta\in\Theta\}$,$\Theta$是参数空间, 待估参数$g(\theta)$是定义在$\Theta$上参数$\theta$的函数, 且

$$g(\theta)=G(\mu_1,\cdots,\mu_k;\nu_2,\ldots,\nu_s).$$

设样本$\widetilde{X}=(X_1,\cdots,X_n)$是从上述分布族中的某分布抽取的简单样本, 将上式中的μ_i和ν_j分别用它们的矩估计量$a_{n,i}$和$m_{n,j}$代替,得

$$\widehat{g(\theta)}=G(a_{n,1},\cdots,a_{n,k};m_{n,2},\cdots,m_{n,s}),$$

则$\widehat{g(\theta)}$作为$g(\theta)$的估计量, 称为$g(\theta)$的矩法估计量(moment estimator). 这种求矩估计量的方法称为矩法(moment method of estimation).

二、分布中未知参数的矩法估计量

对于参数分布族, 设总体分布中含有k个未知参数, $\theta_1, \theta_2, \cdots, \theta_k$,要用矩法估计这$k$个参数的前提是: 总体的前$k$阶矩存在. 那么它的前$k$阶矩(或中心矩)一般是这$k$个参数的函数:

$$\mu_i = g_i(\theta_1, \theta_2, \cdots, \theta_k), \quad i = 1, 2, \cdots, k.$$

假如能解出θ_j:

$$\theta_j = h_j(\mu_1, \mu_2, \cdots, \mu_k), \quad j = 1, 2, \cdots, k.$$

那么用样本矩$a_{n,i}$代替总体矩μ_i,就得到参数θ_j的矩法估计:

$$\widehat{\theta}_j = h_j(a_{n,1}, a_{n,2}, \cdots, a_{n,k}), \quad j = 1, 2, \cdots, k.$$

基本步骤:

(1)写出总体的前k阶矩, 一般是这k个待估参数的函数,
(这一步的目的是为了下一步,有时此步可省)

$$\mu_i = g_i(\theta_1, \theta_2, \cdots, \theta_k), \quad i = 1, 2, \cdots, k;$$

(2)写出待估参数关于总体矩的函数表达式,

$$\theta_j = h_j(\mu_1, \mu_2, \cdots, \mu_k), \quad j = 1, 2, \cdots, k;$$

(3)写出第(2)步中出现的总体矩相应的样本矩;
(4)以相应的样本矩替换第(2)步中出现的总体矩,从而得到参数的矩法估计量,

$$\widehat{\theta}_j = h_j(a_{n,1}, a_{n,2}, \cdots, a_{n,k}), \quad j = 1, 2, \cdots, k.$$

Example

总体X来自指数分布族$\{E(\lambda), \lambda > 0\}$, 设$X_1, X_2, \cdots, X_n$是取自该总体的简单随机样本. 求未知参数$\lambda$的矩法估计,并判断此估计量是否无偏.

解: $\mu_1 = \mathsf{E}X = 1/\lambda$, 即$\lambda = 1/\mathsf{E}X = 1/\mu_1$. 而$a_{n,1} = \overline{X}$, 所以$\lambda$的矩法估计是

$$\widehat{\lambda} = \frac{1}{\overline{X}},$$

但它不是λ的无偏估计. 事实上$\mathsf{E}(\lambda) = gamma(1, \lambda)$. 所以$\sum_{i=1}^n X_i \sim gamma(n, \lambda)$. 从而

$$\mathsf{E}(\widehat{\lambda}) = \mathsf{E}\left(\frac{n}{\sum_{i=1}^n X_i}\right) = \int_0^\infty \frac{n}{y}\frac{\lambda^n}{\Gamma(n)}e^{-\lambda y}y^{n-1}dy = \frac{n}{n-1}\lambda.$$

对$\widehat{\lambda}$略作修改，令

$$\widehat{\lambda}^{\star}=\frac{n-1}{n}\widehat{\lambda}=(n-1)/\sum_{i=1}^{n}X_i,$$

则$\widehat{\lambda}^{\star}$是$\lambda$的无偏估计.

Example

设$X_1, X_2, \cdots, X_n$是来自均匀分布总体$X \sim U(a,b)(a<b)$的一个样本.求参数a与b的矩法估计.

解: $\mu_1 = \mathsf{E}X = (a+b)/2$, $\nu_2 = \mathsf{Var}(X) = (b-a)^2/12$. 解方程组得

$$a = \mu_1 - \sqrt{3\nu_2}, \quad b = \mu_1 + \sqrt{3\nu_2}.$$

用$a_{n,1} = \overline{X}$代替μ_1, 用$m_{n,2} = S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$代替$\nu_2$, 得$a$, b的矩估计量为

$$\widehat{a} = \overline{X} - \sqrt{3}S_n, \quad \widehat{b} = \overline{X} + \sqrt{3}S_n.$$

Example

设$X_1, X_2, \cdots, X_n$是来自二项分布$X \sim B(k, p)$总体的一组样本, $k \geq 1, 0 < p < 1$. 求参数k与p的矩法估计.

解: $\mu_1 = \mathsf{E}X = kp$, $\nu_2 = \mathsf{Var}(X) = kp(1-p)$. 解方程组得

$$p = 1 - \frac{\nu_2}{\mu_1} = \frac{\mu_1 - \nu_2}{\mu_1}$$

和

$$k = \frac{\mu_1}{p} = \frac{\mu_1}{1 - \nu_2/\mu_1} = \frac{\mu_1^2}{\mu_1 - \nu_2}.$$

用$a_{n,1}=\overline{X}$, $m_{n,2}=S_n^2=\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})^2$分别替换$\mu_1$, ν_2,得p及k的矩估计量为

$$\widehat{p}=\frac{\overline{X}-S_n^2}{\overline{X}},$$

$$\widehat{k}=\frac{\overline{X}^2}{\overline{X}-S_n^2}.$$

Example

设$X_1, \cdots, X_n$是从具有成功概率θ的两点分布总体$B(1,\theta)$中抽取的简单随机样本, 其中$0<\theta<1$未知. 求$g(\theta)=\theta(1-\theta)$的矩法估计.

解: 已知总体$X \sim B(1,\theta)$, 则$\mu_1 = E(X) = \theta$, 那么$g(\theta) = \mu_1(1-\mu_1)$. 因此$g(\theta)$的矩法估计量是

$$\widehat{g(\theta)} = \overline{X}(1-\overline{X}).$$

♣ 还有其它思路吗?

矩法估计也可以用于多维总体.

Example

设$\mathbf{X}_1 = (X_{1,1}, \cdots, X_{1,d}), \cdots, \mathbf{X}_n = (X_{n,1}, \cdots, X_{n,d})$是来自多元正态分布总体$\mathbf{X} \sim N(\mu, \mathbf{\Sigma})$ 的一个样本, 其中$\mu = (\mu_1, \mu_2, \cdots, \mu_d)$, $\mathbf{\Sigma}$为$d \times d$的矩阵, 均未知. 求参数μ 和$\mathbf{\Sigma}$ 的估计.

解: 注意到$\mathsf{E}[\mathbf{X}] = \mu$,

$$\mathsf{E}\left[(\mathbf{X} - \mathsf{E}[\mathbf{X}])'(\mathbf{X} - \mathsf{E}[\mathbf{X}])\right] = Var(\mathbf{X}) = \mathbf{\Sigma},$$

它们的估计分别为

$$\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i,$$

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})'(\mathbf{X}_i - \overline{\mathbf{X}}).$$

Example

设$X_1, X_2, \cdots, X_n$是来自Cauchy分布总体X的一个样本, 密度函数为

$$p(x) = \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2}.$$

求参数μ的估计.

三、矩估计量的相合性

Theorem

定理 假如总体的k阶矩存在, 那么$a_{n,k}$是μ_k的(强)相合估计.

Theorem

定理 设$\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k$分别是$\theta_1, \theta_2, \cdots, \theta_k$的(强)相合估计. 假设$g(\theta_1, \theta_2, \cdots, \theta_k)$为参数空间上连续函数, 则$g(\widehat{\theta}_1, \widehat{\theta}_2, \cdots, \widehat{\theta}_k)$是$g(\theta_1, \theta_2, \cdots, \theta_k)$的(强)相合估计.

四、矩估计量的渐近正态性

Theorem

假如总体的$2k$阶矩存在, 那么样本k阶原点矩

$$a_{n,k} \sim AN\big(\mu_k, \frac{\mu_{2k} - \mu_k^2}{n}\big).$$

更一般地

$$(a_{n,1}, \ldots, a_{n,k}) \sim AN\left((\mu_1, \ldots, \mu_k), \frac{1}{n}\mathbf{V}\right)$$

其中$\mathbf{V} = \big(\mu_{i+j} - \mu_i\mu_j\big)_{k\times k}$.

证明: 记$\widetilde{Y}_i = (X_i - \mu_1, X_i^2 - \mu_2, \ldots, X_i^k - \mu_k)$. 则$\widetilde{Y}_i$, $i = 1, 2, \ldots, n$, 是一列均值为0, 协方差矩阵为$\mathbf{V}$ 的i.i.d.随机向量. 由中心极限定理知, 当$n \to \infty$, 有

$$\sqrt{n}\left((a_{n,1}, \ldots, a_{n,k}) - (\mu_1, \ldots, \mu_k)\right) = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \widetilde{Y}_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{Y}_i \xrightarrow{D} N(\widetilde{0}, \mathbf{V}).$$

结论得证.

一般地, 矩法估计量是样本原点矩的函数, 其渐近正态性可由δ-法则得到.

Theorem

(δ-method) 设$\widehat{\theta}_n$是$\theta = (\theta_1, \ldots, \theta_k)$ 的相合渐近正态估计,

$$\widehat{\theta}_n \sim AN(\theta, \mathbf{\Sigma}/k_n),$$

$k_n \to \infty$, $\mathbf{g}(\theta) = (g_1(\theta), \ldots, g_d(\theta))$是$\theta$的可微函数. 则$\mathbf{g}(\widehat{\theta}_n)$是$\mathbf{g}(\theta)$的相合渐近正态估计,

$$\mathbf{g}(\widehat{\theta}_n) \sim AN\left(\mathbf{g}(\theta), \mathbf{H}'\mathbf{\Sigma}\mathbf{H}/k_n\right),$$

其中$\mathbf{H} = \frac{\partial \mathbf{g}}{\partial \theta} = \left(\frac{\partial g_j}{\partial \theta_i}; i, j\right)$ 为$k \times d$的矩阵.

The idea of the proof:

$$\sqrt{k_n}\left(\mathbf{g}(\widehat{\theta}_n)-\mathbf{g}(\theta)\right)\approx\sqrt{k_n}\left(\widehat{\theta}_n-\theta\right)\frac{\partial\mathbf{g}}{\partial\theta}\xrightarrow{D}N(\mathbf{0},\mathbf{H}'\mathbf{\Sigma}\mathbf{H}).$$

证明: 令

$$\mathbf{h}(\mathbf{x})=\begin{cases}\mathbf{0}, & \text{若}\mathbf{x}=\theta;\\ \frac{\mathbf{g}(\mathbf{x})-\mathbf{g}(\theta)-(\mathbf{x}-\theta)\frac{\partial\mathbf{g}}{\partial\theta}}{\|\mathbf{x}-\theta\|} & \text{若}\mathbf{x}\neq\theta.\end{cases}$$

则$\mathbf{h}(\mathbf{x})$在θ处连续, 且

$$\sqrt{k_n}\left(\mathbf{g}(\widehat{\theta}_n)-\mathbf{g}(\theta)\right)=\sqrt{k_n}\left(\widehat{\theta}_n-\theta\right)\frac{\partial\mathbf{g}}{\partial\theta}+\sqrt{k_n}\|\widehat{\theta}_n-\theta\|\mathbf{h}(\widehat{\theta}_n).$$

由

$$\sqrt{k_n}\left(\widehat{\theta}_n-\theta\right)\xrightarrow{d}N(\mathbf{0},\mathbf{\Sigma})$$

得

$$\widehat{\theta}_n\xrightarrow{d}\theta\quad\text{即}\quad\widehat{\theta}_n\xrightarrow{P}\theta.$$

从而$\mathbf{h}(\widehat{\theta}_n)\xrightarrow{P}\mathbf{0}$. 所以

$$\sqrt{k_n}\|\widehat{\theta}_n-\theta\|\mathbf{h}(\widehat{\theta}_n)\xrightarrow{P}\mathbf{0}.$$

另一方面

$$\sqrt{k_n}\left(\widehat{\theta}_n-\theta\right)\frac{\partial\mathbf{g}}{\partial\theta}\xrightarrow{d}N(\mathbf{0},\mathbf{H}'\mathbf{\Sigma}\mathbf{H}).$$

结论得证.

矩法估计的推广

估计方程 如果$g(X;\theta)$是总体X和待估参数θ的函数, 满足

$$E_F g(X;\theta) = \int g(x;\theta)dF(x) = 0.$$

(称为估计方程). 那么将$F(x)$用经验分布函数$F_n(x)$代替, 得到的方程即为

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i;\theta) = \int g(x;\theta)dF_n(x) = 0.$$

上述方程的解$\widehat{\theta}_n$也就是θ的矩法估计.

Example

设总体X的密度函数为$p(x;\theta)$, θ为参数, 未知. 若满足$\frac{d}{d\theta}\int p(x;\theta)dx = \int \frac{\partial}{\partial\theta}p(x;\theta)dx$, 试用估计方程求$\theta$的估计.

由$\int p(x;\theta)dx = 1$得

$$0 = \frac{d}{d\theta}\int p(x;\theta)dx = \int \frac{\partial}{\partial\theta}p(x;\theta)dx = \int \frac{\partial \log p(x;\theta)}{\partial\theta}p(x;\theta)dx.$$

得估计方程

$$E_\theta \frac{\partial \log p(X;\theta)}{\partial\theta} = 0.$$

这样,下述方程的解即为θ的估计:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log p(X_i;\theta)}{\partial\theta} = 0.$$

这是我们后面要讲到的对数似然方程.

矩法估计的前提:

总体有直到k阶的矩(或中心矩)存在(其中,k为未知参数的个数).

矩法估计的优点:

(1)简便且不要求事先知道总体的分布;

(2)一般说来, 矩估计量还是相应参数的相合估计量.

矩法估计的缺点:

(1)在参数分布族场合,没有充分利用已知参数分布族提供的信息;

(2)在小样本场合无突出性质;

(3)不够稳健;

(4)在很多场合下，矩估计量不具有唯一性.

如: 对于泊松分布, $\mathsf{E}X = \mathsf{Var}X = \lambda$, 则

$$\widehat{\lambda} = \overline{X} \quad 和 \quad \widehat{\lambda} = S_n^2$$

均是λ的矩法估计量.

§3.3 估计方法之二——极大似然估计法
(Maximum likelihood estimation)
(仅适用于参数分布族)

似然原理: 一个随机试验的一个结果中最可能出现的是该试验所有可能结果中概率最大的那个结果.

Example

一罐中放着大量的黑球和红球,并假定已知两种球的数目之比是1 : 3,但不知道哪种颜色的球多(即,抽到一个球为黑球的概率或者是1/4或者是3/4).用有放回抽样,从罐中抽取了5只球,结果是:黑,红,黑,黑,黑.试问:罐中的球况比较“像”哪一种? $\hat{p} = 3/4$

解: 从罐中随机取一球, 考察该球的颜色. 设总体为

$$X=\begin{cases}1, & \text{取到的是黑球;}\\ 0, & \text{取到的是红球.}\end{cases}$$

则总体X来自分布族$\{B(1,p):p\in\{1/4,3/4\}\}$. 现抽取了容量为5的一组样本$\widetilde{X}$, 得到样本观测值$\widetilde{x}$为$(1,0,1,1,1)$, 则出现本次观测结果的概率为

$$\mathsf{P}(\widetilde{X}=\widetilde{x})=\mathsf{P}(X_1=1,X_2=0,X_3=1,X_4=1,X_5=1)=(1-p)^1p^4,\text{为参数}p\text{的函数}$$

如果$p=\frac{1}{4}$, 那么得到结果$(1,0,1,1,1)$的概率为$\left(\frac{1}{4}\right)^4\frac{3}{4}$;
如果$p=\frac{3}{4}$, 那么得到结果$(1,0,1,1,1)$的概率为$\left(\frac{3}{4}\right)^4\frac{1}{4}$.
这说明后一种情况比前一种情况更有可能产生目前这一结果. 因此我们有理由认为罐中的球况更有可能是后一种情况, 即$\hat{p}$取$3/4$更合理.

Example

稍微一般地, 设罐子中黑球的所占比例p可能是$\{0, 1/4, 2/4, 3/4, 1\}$中的一种, 但不知道是哪一种. 用有放回抽样,从罐中抽取了5 只球,结果是:黑,红,黑,黑,黑.试问:罐中的球况比较"像"哪一种?

同样考虑p的函数$\mathsf{P}(\widetilde{X} = \widetilde{x}) = (1-p)^1 p^4$, 可得

如果$p = 0$, 那么得到结果(黑,红,黑,黑,黑)的概率为$0^4 \times 1 = 0$;

如果$p = 1/4$, 那么得到结果(黑,红,黑,黑,黑)的概率为$\left(\frac{1}{4}\right)^4 \frac{3}{4}$;

如果$p = 2/4$, 那么得到结果(黑,红,黑,黑,黑)的概率为$\left(\frac{2}{4}\right)^4 \frac{2}{4}$;

如果$p = 3/4$, 那么得到结果(黑,红,黑,黑,黑)的概率为$\left(\frac{3}{4}\right)^4 \frac{1}{4}$;

如果$p = 1$, 那么得到结果(黑,红,黑,黑,黑)的概率为$1^4 \times 0 = 0$.

这说明当$p = 3/4$时最有可能产生(黑, 红, 黑, 黑, 黑) 这一结果. 因此我们有理由认为罐中的球况是$\hat{p} = 3/4$.

Example

更加一般地, 设罐子中黑球的所占比例$p(0 \le p \le 1)$未知. 用有放回抽样,从罐中抽取了5只球,结果是:黑,红,黑,黑,黑.试问:p最有可能取何值?

同样考虑p的函数$\mathsf{P}(\widetilde{X} = \widetilde{x}) = (1-p)^1 p^4$, 当$p$为

$$\arg \sup_{0 \le p \le 1} p^4(1-p) = 4/5$$

时, 最有可能产生(黑, 红, 黑, 黑, 黑) 这一结果. 因此我们有理由认为罐中的球况最有可能是$\hat{p} = 4/5$.

极大似然估计的思想: 设总体含有待估参数θ,它有不止一个的可能取值($\theta \in \Theta$)，我们要在θ的一切可能取值(Θ)之中选出一个使样本观测值出现的概率为最大的θ 值(记为$\widehat{\theta}$)作为θ的估计，并称$\widehat{\theta}$为θ的极大似然估计.

一、似然函数

下面分总体X的分布是离散的与连续的两种情况加以讨论.

Case1: 离散分布场合

一般地, 假设总体分布族$\mathcal{F}=\{F_\theta;\theta\in\Theta\}$ 是离散型的, 其分布列为

$$\mathsf{P}_\theta(X=a_i)=p(a_i;\theta),\qquad i=1,2,\cdots.$$

则样本$\widetilde{X}=(X_1,X_2,\cdots,X_n)$的联合分布列(pmf)为

$$p(\widetilde{x};\theta)=P_\theta(\widetilde{X}=\widetilde{x})=\prod_{i=1}^{n}P_\theta(X=x_i).$$

设观察到样本$\widetilde{X}$的一个观测值$\widetilde{x}$, 那么$p(\widetilde{x};\theta)$就成了θ的函数, 记为$L(\theta;\widetilde{x})$, 如果

$$L(\theta_1;\widetilde{x}) > L(\theta_2;\widetilde{x}),$$

则看上去,这个样本来自总体F_{θ_1}比来自总体F_{θ_2}的可能性更大.

给定$\widetilde{x}$, 作为θ的函数$L(\theta;\widetilde{x})$可看成参数θ对产生观察值$\widetilde{x}$有“多大可能”(likelihood) 的一种度量, 我们称之为似然函数.

Case2: 连续分布场合

当总体分布族$\mathcal{F}=\{F_\theta;\theta\in\Theta\}$为连续型的, 其概率密度函数为$p(x;\theta)$.则样本$\widetilde{X}$的联合密度函数(pdf)为

$$p(\widetilde{x};\theta)=\prod_{i=1}^{n}p(x_i;\theta).$$

设观测到$\widetilde{X}$的一个观测值为$\widetilde{x}$, 那么上式就成了θ的函数,记为$L(\theta;\widetilde{x})$.由于

$$P_\theta\{\widetilde{X}\in Ball(\widetilde{x},\epsilon)\}\approx L(\theta;\widetilde{x})\cdot Vol\{Ball(\widetilde{x},\epsilon)\}.$$

$L(\theta;\widetilde{x})$仍可看成参数θ对产生观察值$\widetilde{x}$有“多大可能”的一种度量.

Definition

定义 设$p(\widetilde{x};\theta)$为样本$\widetilde{X}$的联合密度函数(pdf)或分布列(pmf), $\theta \in \Theta$是未知参数. 对给定的$\widetilde{x}$, 参数θ的函数

$$L(\theta;\widetilde{x}) = p(\widetilde{x};\theta), \quad \theta \in \Theta,$$

称为似然函数(likelihood function). 而$l(\theta;\widetilde{x}) = \log L(\theta;\widetilde{x})$称为对数似然函数(log likelihood function).

二、极大似然估计

Definition

定义 设总体分布族为$\mathcal{F}=\{p(x;\theta);\theta=(\theta_1,\cdots,\theta_k)\in\Theta\}$是一个参数分布族, $p(x;\theta)$为密度函数(pdf)或分布列(pmf). $\widetilde{x}=(x_1,\cdots,x_n)$是由$\mathcal{F}$中的一个总体$p(x;\theta)$产生的简单随机样本$\widetilde{X}=(X_1,\cdots,X_n)$的观测值. 这时似然函数为

$$L(\theta;\widetilde{x})=\prod_{i=1}^{n}p(x_i;\theta),\quad \theta\in\Theta.$$

假如存在统计量$\widehat{\theta}(\widetilde{X})=\widehat{\theta}(X_1,\cdots,X_n)$, 使得

$$L(\widehat{\theta}(\widetilde{x});\widetilde{x})=\sup_{\theta\in\Theta}L(\theta;\widetilde{x}).$$

Definition

(接上页)

或等价使得

$$l(\widehat{\theta}(\widetilde{x});(\widetilde{x})) = \sup_{\theta\in\Theta} l(\theta;\widetilde{x}).$$

则称$\widehat{\theta}(\widetilde{X})$为$\theta$的极大似然估计量(Maximum likelihood estimator). 简记为MLE.

MLE的基本思想: 用“最像”θ的统计量去估计θ.

求MLE的基本步骤:

(1)写出似然函数$L(\theta;\widetilde{x}),\theta\in\Theta$;

(2)求$\widehat{\theta}$,使得

$$L(\widehat{\theta}(\widetilde{x});\widetilde{x})=\sup_{\theta\in\Theta}L(\theta;\widetilde{x}).$$

当$L(\theta;\widetilde{x})=p(\widetilde{x};\theta)$ 前乘上一个不依赖于θ的正数(可以依赖于$\widetilde{x}$)时, 它的最大值点不改变. 常常把与$p(\widetilde{x};\theta)$成比例的函数也称为似然函数.

当似然函数关于参数可微时, 求MLE常常转变为求下列方程的解:

$$\frac{\partial}{\partial\theta_i}L(\theta_1,\theta_2,\cdots,\theta_k;\widetilde{x})=0,\quad i=1,2,\cdots,k,$$

或

$$\frac{\partial}{\partial\theta_i}l(\theta_1,\theta_2,\cdots,\theta_k;\widetilde{x})=0,\quad i=1,2,\cdots,k,$$

前者称为似然方程,后者称为对数似然方程.

Example

设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$的一个样本, 求未知参数μ和σ^2的极大似然估计.

解: 似然函数为

$$L(\mu, \sigma^2; \widetilde{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\},$$
$$-\infty < \mu < \infty, \sigma > 0.$$

对数似然函数为

$$l(\mu, \sigma^2; \widetilde{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$
$$-\infty < \mu < \infty, \sigma > 0.$$

则对数似然方程为

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0, \\ \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0, \end{cases}$$

解方程, 得

$$\begin{cases} \widehat{\mu}(\widetilde{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}, \\ \widehat{\sigma^2}(\widetilde{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = s_n^2. \end{cases}$$

经微分法检验知

$$\widehat{\mu}(\widetilde{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X},$$

$$\widehat{\sigma^2}(\widetilde{X}) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 = S_n^2,$$

是参数μ, σ^2的MLE, 即$(\overline{X}, S_n^2)$是参数(μ, σ^2)的MLE.

用微分法求解MLE的基本步骤:

(1)写出似然函数$L(\theta;\widetilde{x}),\theta\in\Theta$;
(有时需要写出对数似然函数$l(\theta;\widetilde{x})=\log L(\theta;\widetilde{x}),\theta\in\Theta$;)

(2)列出似然方程$\frac{\partial}{\partial\theta}L(\theta;\widetilde{x})=0$;
(有时需要列出对数似然方程$\frac{\partial}{\partial\theta}L(\theta;\widetilde{x})=0$;)

(3)解(对数)似然方程, 得$\theta=h(\widetilde{x})$;

(4)经微分法检验, 得θ的极大似然估计量为$\widehat{\theta}=h(\widetilde{X})$.

Example

设总体X的概率分布列为$\begin{pmatrix} 1 & 2 & 3 \\ \theta & \frac{\theta}{2} & 1-\frac{3\theta}{2} \end{pmatrix}$, 其中$0<\theta<\frac{2}{3}$未知. 现得到样本观察值$(2,3,2,1,3)$, 求$\theta$的矩估计值和极大似然估计值.

解: (矩估计) 由于

$$\mu_1 = \mathsf{E}(X) = \theta + 2\times\frac{\theta}{2} + 3\times\left(1-\frac{3\theta}{2}\right) = 3-\frac{5\theta}{2} \implies \theta = \frac{2}{5}(3-\mu_1).$$

故θ的矩估计量为$\widehat{\theta} = \frac{2}{5}(3-\overline{X})$. 现有$\overline{x}=2.2$, 所以$\theta$的矩估计值为

$$\widehat{\theta} = \frac{2}{5}(3-\overline{x}) = 0.32.$$

(极大似然估计) 似然函数为

$$L(\theta)=\mathsf{P}\{\widetilde{X}=\widetilde{x}\}=\frac{\theta}{2}\cdot\left(1-\frac{3\theta}{2}\right)\cdot\frac{\theta}{2}\cdot\theta\cdot\left(1-\frac{3\theta}{2}\right)=\frac{1}{16}\theta^3(2-3\theta)^2.$$

对数似然函数为

$$\ln L(\theta)=-\ln 16+3\ln\theta+2\ln(2-3\theta).$$

由$\dfrac{d\ln L(\theta)}{d\theta}=\dfrac{3}{\theta}-\dfrac{6}{2-3\theta}=0$得解为$\theta=0.4$. 经微分法检验, 可得$\theta$的极大似然估计值为

$$\widehat{\theta}_{\mathrm{MLE}}=0.4.$$

Example

设$X_1, X_2, \cdots, X_n$是取自指数分布的总体$E(\lambda)$的一个样本, 求未知参数$\lambda(\lambda > 0)$的极大似然估计量.

解: 似然函数为

$$L(\lambda; \widetilde{x}) = \lambda^n \exp\{-\lambda \sum_{i=1}^{n} x_i\}, \quad \lambda > 0, x_i > 0, i = 1, 2, \cdots, n.$$

对数似然函数为　$l(\lambda; \widetilde{x}) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$. 则对数似然方程为

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

解方程, 得　$\widehat{\lambda}(\widetilde{x}) = n / \sum_{i=1}^{n} x_i$. 经微分法检验知$\lambda$的极大似然估计为

$$\widehat{\lambda}(\widetilde{X}) = n / \sum_{i=1}^{n} X_i = \frac{1}{\overline{X}}.$$

Example

设某厂生产的灯泡的寿命服从指数分布$E(\lambda)(\lambda > 0)$. 现抽取n只灯泡做试验, 试验到第$r(\leq n)$只灯泡失效时停止试验, 记录这r只灯泡的试验时间为$X_{(1)}, X_{(2)}, \cdots, X_{(r)}$. 求$\lambda$的MLE.

定时截尾(寿命)试验

设$X_1, X_2, \cdots, X_n$是这n个灯泡的寿命, 则$(X_1, X_2, \cdots, X_n)$即为一个样本, 这是一个完全样本.
而现在我们只观察到这一样本的次序统计量$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$的前$r$个, $X_{(1)}, X_{(2)}, \cdots, X_{(r)}$. 当$r < n$时, $(X_{(1)}, X_{(2)}, \cdots, X_{(r)})$是个不完全样本.

不完全样本的其它形式: truncated (截尾) data,　censored (删失) data, contaminated (污染) data,　missing (缺失) data .....

略解: 总体的密度函数为 $p(x)=\lambda e^{-\lambda x}, x>0$. 似然函数(与$(X_{(1)},\cdots,X_{(r)})$ 的联合密度函数形式相同)为

$$\begin{aligned}L(\lambda;x_{(1)},\cdots,x_{(r)}) &= \frac{n!}{(n-r)!}p(x_{(1)})\cdots p(x_{(r)})[1-F(x_{(r)})]^{n-r}\\ &=\frac{n!}{(n-r)!}\lambda^r\exp\left\{-\lambda\left[x_{(1)}+\cdots+x_{(r)}+(n-r)x_{(r)}\right]\right\},\\ &\quad \lambda>0, 0<x_{(1)}\le\cdots\le x_{(r)}.\end{aligned}$$

令

$$\frac{\partial l}{\partial\lambda}=\frac{\partial\log L}{\partial\lambda}=\frac{r}{\lambda}-\{x_{(1)}+\cdots+x_{(r)}+(n-r)x_{(r)}\}=0$$

解得$\widehat{\lambda}=r/\{x_{(1)}+\cdots+x_{(r)}+(n-r)x_{(r)}\}$. 经微分法检验，得到$\lambda$的MLE为

$$\widehat{\lambda}=r/\{X_{(1)}+\cdots+X_{(r)}+(n-r)X_{(r)}\}.$$

Example

设$(X_1, X_2, \cdots, X_n)$是取自均匀分布$U(0,\theta)$的一个样本, 其中$\theta > 0$.
求θ的MLE.

解: 似然函数为

$$L(\theta) = p(\widetilde{x};\theta) = \begin{cases} \frac{1}{\theta^n}, & 0 \le x_1, \ldots, x_n \le \theta; \\ 0, & \text{其它}. \end{cases}$$

故对数似然函数为

$$l(\theta) = \log p(\widetilde{x};\theta) = \begin{cases} -n\log\theta, & 0 \le x_1, \ldots, x_n \le \theta; \\ 0, & \text{其它}. \end{cases}$$

由于$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} \neq 0$, 所以这里不能用微分法来求$\widehat{\theta}_{MLE}$.

下面用定义法来求$\widehat{\theta}_{MLE}$.

由于$L(\theta)>0$需要$0\leq x_i\leq\theta,\quad i=1,2,\ldots,n$, 即此时$\theta$的取值范围为$0\leq x_{(n)}\leq\theta$, 其中$x_{(n)}=\max\{x_1,x_2,\cdots,x_n\}$. 注意到$L(\theta)=\frac{1}{\theta^n}$对$\theta\geq x_{(n)}$的$\theta$是一个严格单调的减函数, θ越小, $L(\theta)$的值越大, 故有

$$\arg\sup_{\theta\geq x_{(n)}} L(\theta)=x_{(n)}.$$

所以θ的MLE为$\widehat{\theta}_{MLE}=X_{(n)}$.

注意: 当似然函数的非零区域与未知参数有关时,通常无法通过求解似然方程来获得参数的极大似然估计, 这时一般可从定义出发直接来求似然函数的极大值点(定义法).

Example

(续)讨论MLE $T(\widetilde{X}) = X_{(n)}$的性质.

解: (1) 无偏性. T的概率密度函数为

$$g(t;\theta) = \begin{cases} \frac{nt^{n-1}}{\theta^n}, & 0 \le t \le \theta; \\ 0, & \text{其它}. \end{cases}.$$

故有

$$\mathsf{E}X_{(n)} = \int_0^\theta t \cdot \frac{n}{\theta^n} t^{n-1} dt = \frac{n}{n+1}\theta.$$

所以$X_{(n)}$ 不是θ的无偏估计, 而是渐近无偏估计.

$\widehat{\theta}_1 = (1 + \frac{1}{n})X_{(n)} = \frac{n+1}{n}X_{(n)}$是$\theta$的无偏估计.

另, 由于$\mu_1 = \mathsf{E}X = \frac{\theta}{2}$, 则$\theta = 2\mu_1$. 所以$\theta$的矩法估计为

$$\widehat{\theta}_2 = 2\overline{X}.$$

而

$$\mathsf{E}\widehat{\theta}_2 = 2\mathsf{E}\overline{X} = 2 \cdot \frac{0+\theta}{2} = \theta,$$

故$\widehat{\theta}_2$为θ的无偏估计量.

(2)比较$\widehat{\theta}_1$和$\widehat{\theta}_2$这两个无偏估计量的有效性.

先来求$\widehat{\theta}_1$的方差,

$$\mathsf{E}X_{(n)}^2 = \int_0^\theta t^2 \cdot \frac{n}{\theta^n} t^{n-1} dt = \frac{n}{n+2}\theta^2.$$

所以

$$\mathsf{Var}(\widehat{\theta}_1) = \mathsf{E}\widehat{\theta}_1^2 - \theta^2 = (1+\frac{1}{n})^2 \frac{n}{n+2}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

再来看看$\widehat{\theta}_2$的方差.

$$\mathsf{Var}(\widehat{\theta}_2) = 4 \cdot \mathsf{Var}(\overline{X}) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

可见当$n \geq 2$时, $\widehat{\theta}_1$比矩法估计量$\widehat{\theta}_2$有效. $n = 1$时, 两者一致.

(3)比较$\widehat{\theta}_{MLE}=X_{(n)}$和矩法估计量$\widehat{\theta}_2=2\overline{X}$这两个估计量的均方误差.
MLE $X_{(n)}$的均方误差为

$$\begin{aligned}MSE(\widehat{\theta}_{MLE}) =&E(X_{(n)}-\theta)^2=EX_{(n)}^2-2\theta EX_{(n)}+\theta^2\\ =&\theta^2\left(\frac{n}{n+2}-2\frac{n}{n+1}+1\right)\\ =&\frac{2\theta^2}{(n+2)(n+1)}\\ <&MSE(\widehat{\theta}_2)=\mathsf{Var}(\widehat{\theta}_2)=\frac{\theta^2}{3n};n\geq 2.\end{aligned}$$

$n=1$时, 两者一致.

(4) 相合性. 对任意的$0 < \epsilon < \theta$;

$$\begin{aligned} P(|X_{(n)} - \theta| \geq \epsilon) =& P(X_{(n)} \geq \theta + \epsilon) + P(X_{(n)} \leq \theta - \epsilon) \\ =& 0 + \big(P(X \leq \theta - \epsilon)\big)^n \\ =& \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0, \quad \text{当} \quad n \to \infty. \end{aligned}$$

故$X_{(n)}$是θ 的相合估计.

(或用“Chebyshev不等式”来得到相合性)

注: 极大似然估计也不一定是唯一的.(见书本例3.3.7)

Example

设$(X_1, X_2, \cdots, X_n)$是取自均匀分布$U(\theta, \theta+1)$的一个样本, 其中$\theta \in \mathbb{R}$未知. 求θ的MLE.

解: 似然函数为

$$L(\theta) = p(\widetilde{x}; \theta) = \begin{cases} 1, & \theta \le x_{(1)} \le x_{(n)} \le \theta + 1; \\ 0, & \text{其它}. \end{cases} = \begin{cases} 1, & x_{(n)} - 1 \le \theta \le x_{(1)}; \\ 0, & \text{其它}. \end{cases}$$

这时, 似然函数只有0和1两个值. 可见只要θ满足$x_{(n)} - 1 \le \theta \le x_{(1)}$都可使似然函数达到其极大值1. 如: $\widehat{\theta}_1 = X_{(1)}$, $\widehat{\theta}_2 = X_{(n)} - 1$都是$\theta$的MLE. 事实上, 对于任给的$0 \le \lambda \le 1$, $\widehat{\theta} = \lambda X_{(1)} + (1-\lambda)(X_{(n)} - 1)$都是$\theta$的MLE.

极大似然估计的性质:

(I) 极大似然估计的不变原则:

若$\eta = g(\theta)$ 是一一变换, 则

$$\widehat{\theta} \text{ 是}\theta \text{ 的}MLE \implies \widehat{\eta} = g(\widehat{\theta}) \text{ 是}\eta = g(\theta) \text{ 的}MLE.$$

因为参数η的似然函数为

$$L^*(\eta; \widetilde{x}) = L^*(g(\theta); \widetilde{x}) = L(\theta; \widetilde{x}) = L(g^{-1}(\eta); \widetilde{x}).$$

从而

$$\sup_{\eta} L^*(\eta; \widetilde{x}) = \sup_{\eta} L(g^{-1}(\eta); \widetilde{x}) = \sup_{\theta} L(\theta; \widetilde{x}) = L(\widehat{\theta}; \widetilde{x}) = L^*(g(\widehat{\theta}); \widetilde{x}).$$

所以$g(\widehat{\theta}) = \arg\sup_{\eta} L^*(\eta; \widetilde{x})$, 即$\widehat{\eta} = g(\widehat{\theta})$.

Theorem

定理 设$\widehat{\theta}_{MLE}$是θ的MLE, 则对任意一可测函数$g(\theta)$, $g(\widehat{\theta}_{MLE})$是$g(\theta)$的$MLE$.

(II) 极大似然估计是充分统计量的函数

Theorem

设$T(\widetilde{X})$是θ的充分统计量, $\widehat{\theta}_{MLE}$是θ的MLE, 则$\widehat{\theta}_{MLE}$是$T(\widetilde{X})$的函数.

证明: 设样本$\widetilde{X} = (X_1, \ldots, X_n)$的联合pdf 或pmf 为$p(\widetilde{x}; \theta)$, 由因子分解定理知

$$L(\theta; \widetilde{x}) = p(\widetilde{x}; \theta) = g(T(\widetilde{x}), \theta)h(\widetilde{x}).$$

所以

$$\widehat{\theta}_{MLE}(\widetilde{x}) = \arg\sup_{\theta} L(\theta) = \arg\sup_{\theta} g(T(\widetilde{x}), \theta)$$

是$T(\widetilde{x})$的函数.

(III) 极大似然估计的渐近正态性(仅考虑θ为一维的情形)

Theorem

定理设$\mathcal{F}=\{p(x;\theta):\theta\in\Theta\}$是一个概率密度或分布列族, Θ为直线上的非退化区间, $X_1,X_2,\cdots,X_n$是从$\mathcal{F}$中某个总体$X\sim p(x;\theta_0)$产生的简单随机样本. 当$p(x;\theta)$满足适当正则条件*(regularity condition)*时, 则在参数θ的未知真值θ_0为Θ的一个内点的情况下,其似然方程有一个解,记为$\widehat{\theta}_n$ 满足: $\widehat{\theta}_n$依概率收敛于真值θ_0, 且

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)\xrightarrow{d}N\left(0,\frac{1}{I(\theta_0)}\right),\quad \text{当}\quad n\to\infty.$$

其中

$$I(\theta)=E_\theta\left[\left(\frac{\partial\log p(X;\theta)}{\partial\theta}\right)^2\right].$$

正则条件:

(1) 对一切$\theta \in \Theta$, 偏导数

$$\frac{\partial \log p}{\partial \theta}, \quad \frac{\partial^2 \log p}{\partial \theta^2}, \quad \frac{\partial^3 \log p}{\partial \theta^3} \quad \text{存在.}$$

(2) 对一切$\theta^* \in \Theta$, 存在θ^*的一个邻域U_{θ^*}和函数$F_1(x)$, $F_2(x)$, $H(x)$使得

$$\left|\frac{\partial p}{\partial \theta}\right| \le F_1(x), \quad \left|\frac{\partial^2 p}{\partial \theta^2}\right| \le F_2(x), \quad \left|\frac{\partial^3 \log p}{\partial \theta^3}\right| \le H(x),$$

对$\theta \in U_{\theta^*}$成立. 其中F_1, F_2在实轴上可积, 而$H(x)$满足

$$\int H(x)p(x;\theta)dx \le M, \quad \forall \theta \in U_{\theta^*}.$$

(3) 对一切$\theta \in \Theta$, 有

$$0 < I(\theta) = \mathsf{E}\left[\frac{\partial \log p(X;\theta)}{\partial \theta}\right]^2 < \infty.$$

主要证明思路: 对数似然函数为$l(\theta;\widetilde{x})=\sum_{i=1}^{n}\log p(x_i;\theta)$. 由Taylor 公式, 可得

$$l'(\theta;\widetilde{X})=l'(\theta_0;\widetilde{X})+(\theta-\theta_0)l''(\theta_0;\widetilde{X})+remainder.$$

现忽略余项(详细证明中需讨论), 并将θ用方程$l'(\theta;\widetilde{X})=0$的解$\widehat{\theta}_n$代入得

$$0\doteq l'(\theta_0;\widetilde{X})+(\widehat{\theta}_n-\theta_0)l''(\theta_0;\widetilde{X}).$$

即

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)\doteq\frac{\frac{1}{\sqrt{n}}l'(\theta_0;\widetilde{X})}{-\frac{1}{n}l''(\theta_0;\widetilde{X})}.$$

对于分子,

$$l'(\theta;\widetilde{X}) = \sum_{i=1}^{n} \frac{\partial \log p(X_i;\theta)}{\partial \theta} =: \sum_{i=1}^{n} Y_i$$

为i.i.d. 随机变量之和, 且

$$\begin{aligned} \mathsf{E}Y_i =& \mathsf{E}_\theta \left[\frac{\partial \log p(X;\theta)}{\partial \theta} \right] = \int \frac{1}{p(x;\theta)} \frac{\partial p(x;\theta)}{\partial \theta} p(x;\theta) dx \\ =& \int \frac{\partial p(x;\theta)}{\partial \theta} dx = \frac{d}{d\theta} \int p(x;\theta) dx = 0, \end{aligned}$$

$$\mathsf{E}Y_i^2 = \mathsf{E}_\theta \left[\left(\frac{\partial \log p(X;\theta)}{\partial \theta} \right)^2 \right] = I(\theta).$$

由中心极限定理

$$\frac{1}{\sqrt{n}} l'(\theta;\widetilde{X}) \xrightarrow{D} N(0, I(\theta)), \quad \text{当} \quad n \to \infty.$$

对于分母,

$$l''(\theta;\widetilde{X})=\sum_{i=1}^{n}\frac{\partial^2\log p(X_i;\theta)}{\partial\theta^2}=:\sum_{i=1}^{n}Z_i$$

也是i.i.d.随机变量之和, 且

$$\begin{aligned}\mathsf{E}Z_i=&\mathsf{E}_\theta\left[\frac{\partial^2\log p(X;\theta)}{\partial\theta^2}\right]=\int\frac{\partial}{\partial\theta}\left[\frac{1}{p(x;\theta)}\frac{\partial p(x;\theta)}{\partial\theta}\right]p(x;\theta)dx\\=&\int\frac{\partial^2 p(x;\theta)}{\partial\theta^2}dx+\int-\left[\frac{1}{p(x;\theta)}\frac{\partial p(x;\theta)}{\partial\theta}\right]^2p(x;\theta)dx\\=&\frac{d^2}{d\theta^2}\int p(x;\theta)dx-\mathsf{E}_\theta\left[\left(\frac{\partial\log p(X;\theta)}{\partial\theta}\right)^2\right]=-I(\theta).\end{aligned}$$

由大数律, 可知

$$-\frac{1}{n}l''(\theta;\widetilde{X})\xrightarrow{P}I(\theta),\quad\text{当}\quad n\to\infty.$$

因此

$$\frac{\frac{1}{\sqrt{n}}l'(\theta;\widetilde{X})}{-\frac{1}{n}l''(\theta;\widetilde{X})} \xrightarrow{D} N\left(0,\frac{1}{I(\theta)}\right), \quad \text{当} \quad n\to\infty.$$

从而

$$\frac{\frac{1}{\sqrt{n}}l'(\theta_0;\widetilde{X})}{-\frac{1}{n}l''(\theta_0;\widetilde{X})} \xrightarrow{D} N\left(0,\frac{1}{I(\theta_0)}\right), \quad \text{当} \quad n\to\infty.$$

Corollary

推论设总体分布族是单参数指数型的

$$p(x;\theta) = c(\theta)\exp\{Q_1(\theta)T(x)\}h(x),$$

其中$c(\theta)$和$Q_1(\theta)$有一至三阶连续的导数. 则定理的结论成立.

估计方法之三——最小二乘法 (Method of least squares)

考虑如下模型: 设样本$Y_1, Y_2, \cdots, Y_n$满足:

$$Y_i = \mu_i(\theta) + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{1}$$

观察值 = 理论值 + 误差.

其中θ为未知参数, $\mu_i(\cdot)$为已知函数, $\epsilon_1, \cdots, \epsilon_n$为i.i.d.随机误差, 期望为0, 方差为$\sigma^2$. 我们目的是估计$\theta$, 进而估计$\mu_i(\theta)$.

记

$$Q(\theta)=\sum_{i=1}^{n}(Y_i-\mu_i(\theta))^2,$$

最小二乘法就是用满足

$$Q(\widehat{\theta})=\min_{\theta}Q(\theta)\text{ 即 }\widehat{\theta}=\arg\min_{\theta}Q(\theta),$$

的$\widehat{\theta}$来估计, $\widehat{\theta}$称为θ的最小二乘法估计, 记为LSE.

Example

设样本$X_1, X_2, \cdots, X_n$是来自总体X的i.i.d.样本, X的期望存在, 求$\mathsf{E}X$的LSE.

解: 记$\theta = \mathsf{E}X$和$\epsilon_i = X_i - \mathsf{E}X_i$, 则

$$X_i = \theta + \epsilon_i,$$

即为模型(1), θ的LSE为

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (X_i - \theta)^2.$$

易得$\widehat{\theta} = \frac{1}{n}\sum_i X_i = \overline{X}$.

注:

$$\mathsf{E}X = \arg\min_{\theta} \mathsf{E}(X - \theta)^2.$$

线性回归模型

当$\mu_i(\theta)$为参数θ的线性函数时, 模型(1)称为线性模型, 例如

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

其中$\theta = (\beta_0, \beta_1)$为未知参数.

求LSE, 即求

$$Q(\theta) = \sum_{i=1}^{n} (Y_i - \beta_0 - x_i\beta_1)^2$$

的最小值点,这时只要解方程组:

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0, \\
\frac{\partial Q}{\partial \beta_1} &= -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i = 0.
\end{aligned}$$

由此方程组可解得LSE为

$$\begin{aligned}\widehat{\beta}_0 =& \overline{Y} - \overline{x}\widehat{\beta}_1,\\ \widehat{\beta}_1 =& S_{xx}^{-1} S_{xy},\end{aligned}$$

其中

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}).$$

有了β_0和β_1的估计后, 称

$$\widehat{\epsilon}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, \quad i = 1, 2, \cdots, n$$

为残差(residuals), 它们的平方和$RSS = \sum_{i=1}^{n} \widehat{\epsilon}_i^2$ 称为残差平方和(residual sum of squares). 可用$\widehat{\sigma^2} = RSS/n$估计$\sigma^2$.

最小一乘法(Method of least abosulte deviation)

设样本$Y_1, Y_2, \cdots, Y_n$满足:

$$Y_i = \mu_i(\theta) + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{2}$$

$\epsilon_1, \cdots, \epsilon_n$为i.i.d.随机误差, 分布对称.

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |Y_i - \mu_i(\theta)|.$$

例1: 当$\theta = \mu_i(\theta)$时,

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} |Y_i - \theta| = \text{sample median}.$$

例2: 当$\mu_i(\theta)$为参数$\theta(\beta_0, \beta_1)$的线性函数时,

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 x_i|.$$

M 估计

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho\big(Y_i - \mu_i(\theta)\big).$$

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho\big(Y_i - \beta_0 - \beta_1 x_i\big),$$

其中$\rho(\cdot)$称为损失函数(Loss function).

在Quantile regression中, 常定义Loss function为

$$\rho_\tau(y) = y(\tau - I\{y < 0\}), \quad \text{其中} \quad 0 < \tau < 1.$$