第一节 参数估计的方法

一、最大似然估计法（Maximum Likelihood Estimate，简记为MLE）

● 似然函数：称

$$L(x_1,\cdots,x_n;\ \theta) = \prod_{i=1}^{n} f(x_i,\ \theta)$$

为参数θ的，关于样本$x_1,\cdots,x_n$的似然函数。

概率论中，θ不变，$x_1,\cdots,x_n$是变元，称联合分布；

数理统计中，$x_1,\cdots,x_n$已知（不变），θ未知，是函数变元。

● 称似然函数的（一个）最大值点（若存在）$\widehat{\theta}_n = \widehat{\theta} = \varphi(x_1,\cdots,x_n)$为$\theta$的最大似然估计。

第一节 参数估计的方法

- 经典统计中最重要的估计方法，首先想到。

- 具有许多好的性质，某些标准下最优。

- 思想：用最可能（most likely）产生数据的参数值作为估计值。

　　例如：射击10枪，7中，$X_1, \cdots, X_{10}$独立同分布，共同分布为$B(1, p)$。若只许你猜p为0.2或0.8，如何猜；若允许你在$(0, 1)$中猜，如何猜?

- 求解方法：一般为数学上求最大值方法，取对数（乘积式变为和式），求导，令其等于0。

- 如解不唯一，任何一个均为MLE。

第一节 参数估计的方法

- 强相合性（strong consistency）：

$$P\left(\lim_{n\to\infty}\hat{\theta}_n=\theta\right)=1 \qquad \text{（实变中的几乎处处收敛）}$$

- （弱）相合性（consistency）：$\forall\varepsilon>0$，

$$\lim_{n\to\infty}P\left(\|\hat{\theta}_n-\theta\|<\varepsilon\right)=1 \qquad \text{（实变中的依测度收敛）}$$

- 理论上，强相合性可推出弱相合性，反之不成立。应用中，一般无区别（一般说相合性，指的是弱相合性）。

- 直观意义：随样本量增大，估计值可任意接近目标。

- 区别原因：随机变量的收敛比一般收敛复杂。

第一节 参数估计的方法

几种常见分布:

1． 两点（Bernoulli）分布， $X_i \sim B(1,p)$， $x_1,\cdots,x_n$为数据，取值0、1，

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}$$

$$L(x_1,\cdots,x_n,p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

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

$$log\big(L(p)\big) = \sum_{i=1}^{n} x_i log(p) + \left(n - \sum_{i=1}^{n} x_i\right) log(1-p)$$

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

唯一解为$\hat{p} = \hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} x_i =: \overline{x} = \frac{v}{n}$，概率论中常说的频率。

相合性：由强大数律，$\lim\limits_{n\to\infty} \hat{p}_n = p$ (a.s.)，即频率→概率。

第一节 参数估计的方法

2. 指数分布，$X_i \sim Exp(\lambda)$，

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

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$log\big(L(\lambda)\big) = nlog\lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial log(L(\lambda))}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

唯一解为$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = {}^{1}/_{\bar{x}}$。

相合性：$\bar{x} \to E(X) = 1/\lambda$，故 $\hat{\lambda} \to \lambda$ (a.s.)

$1/\lambda$的直观意义：平均寿命。

第一节 参数估计的方法

3. 正态分布，$X_i \sim N(\mu, \sigma^2)$，令$\delta = \sigma^2$，

$$L(\mu, \delta) = \left(\frac{1}{\sqrt{2\pi\delta}}\right)^n e^{-\frac{1}{2\delta}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$log\big(L(\mu, \delta)\big) = nlog(C) - \frac{n}{2}log(\delta) - \frac{1}{2\delta}\sum_{i=1}^{n}(x_i - \mu)^2$$

其中C为某常数，

似然方程组为：

$$\begin{cases} \dfrac{\partial logL}{\partial \mu} = \dfrac{1}{\delta}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \dfrac{\partial logL}{\partial \delta} = -\dfrac{n}{2\delta} + \dfrac{1}{2\delta^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

唯一解为：

$$\begin{cases} \hat{\mu} =: \dfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \qquad \text{（样本均值）} \\ \hat{\sigma}^2 =: \hat{\delta} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{cases}$$

第一节 参数估计的方法

相合性:

由大数率， $P\left(\lim\limits_{n\to\infty}\hat{\mu}=\mu\right)=1$;

又由 $P\left(\lim\limits_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i^2=EX^2\right)=1$,

故而

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i^2-\lim_{n\to\infty}\hat{\mu}^2=EX^2-(EX)^2\right)=1,$$

即 $P\left(\lim\limits_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=var(X)\right)=1$,

即 $P\left(\lim\limits_{n\to\infty}\hat{\delta}=\delta\right)=1$。

注意其中有: $\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$

这个常用等式。

第一节 参数估计的方法

4. 威布尔（Weibull）分布

$$F(x,m,\eta)=1-exp(-(\frac{x}{\eta})^m)$$ （当$x>0$，参数$m>0$，$\eta>0$。）

$$\Rightarrow \quad f(x,m,\eta)=\frac{m}{\eta^m}x^{m-1}e^{-(\frac{x}{\eta})^m}$$

故 $$L(m,\eta)=\frac{m^n}{\eta^{mn}}\left(\prod_{i=1}^n x_i\right)^{m-1}e^{-\frac{\sum_{i=1}^n x_i^m}{\eta^m}}。$$

取对数，求偏导，可得似然方程组。计算机求解。

无显式解，如何证明相合性?

第一节 参数估计的方法

5. 均匀分布 $X \sim U[a,b]$ **（或简化版本** $X \sim U[0,\theta]$**）**

$$f(x,a,b)=\begin{cases}\dfrac{1}{b-a} & 若 \quad a\le x\le b\\ 0 & 否则\end{cases}$$

$$L(a,b)=\left(\frac{1}{b-a}\right)^n\prod_{i=1}^{n} I_{[a,b]}(x_i),$$

$$所以\ L(a,b)=\begin{cases}\left(\dfrac{1}{b-a}\right)^n & 当 \quad a\le \min\limits_{1\le i\le n} x_i \ 且\ b\ge \max\limits_{1\le i\le n} x_i\\ 0 & 否则\end{cases}$$

故 $\hat{a}=\min\limits_{1\le i\le n} x_i$，$\hat{b}=\max\limits_{1\le i\le n} x_i$ **（不连续，不可导）。**

相合性：教材中有证明，用到实变函数方法。

第一节 参数估计的方法

二、矩估计法（Moment Estimate）

设$\theta=(\theta_1,\cdots,\theta_m)$，$X$的$k$阶矩存在有限，（理论）值为

$$\begin{cases} V_1=E\left(X^1\right)=g_1(\theta_1,\cdots,\theta_m) \\ \qquad\qquad\cdots\qquad\cdots \\ V_m=E(X^m)=g_m(\theta_1,\cdots,\theta_m) \end{cases}$$

样本矩为

$$\begin{cases} \tilde{V}_1=\dfrac{1}{n}\sum_{i=1}^{n}x_i \\ \qquad\cdots\qquad\cdots \\ \tilde{V}_m=\dfrac{1}{n}\sum_{i=1}^{n}x_i^m \end{cases}$$

由大数律（一定条件下），$\tilde{V}_k\to V_k$（当样本量$\to\infty$），令其相等，得到估计方程组

第一节 参数估计的方法

$$
\begin{cases}
g_1(\theta_1,\cdots,\theta_m)=\tilde{V}_1 \\
\qquad\cdots\qquad\cdots \\
g_m(\theta_1,\cdots,\theta_m)=\tilde{V}_m
\end{cases}
$$

若解为

$$
\begin{cases}
\tilde{\theta}_1=f_1(\tilde{V}_1,\cdots,\tilde{V}_m) \\
\qquad\cdots\qquad\cdots \\
\tilde{\theta}_m=f_m(\tilde{V}_1,\cdots,\tilde{V}_m)
\end{cases}
$$

则称其为$(\theta_1,\cdots,\theta_m)$的矩估计。（存在性、唯一性）

矩估计历史上曾有重要地位（Pearson），后Fisher力推MLE。

有时MLE收敛速度更快。

第一节 参数估计的方法

几种常见分布：

1. 两点（Bernoulli）分布， $X_i \sim B(1,p)$，

此时$m=1$，而$V_1=EX=p$，$\tilde{V}_1=\bar{x}$，故$\tilde{p}=\bar{x}$，与MLE相同。

2. 指数分布，$X_i \sim Exp(\lambda)$，

此时$V_1=EX={}^1/_{\lambda}$，$\tilde{V}_1=\bar{x}$，故$\tilde{\lambda}={}^1/_{\bar{x}}$，与MLE一致。

3. 正态分布，$X_i \sim N(\mu,\sigma^2)$

此时$m=2$，$V_1=EX=\mu$，$\tilde{V}_1=\bar{x}$，

$V_2=EX^2=\sigma^2+\mu^2$，$\tilde{V}_2=\frac{1}{n}\sum_{i=1}^{n} x_i^2$，

第一节 参数估计的方法

解方程易知，$(\tilde{\mu}, \tilde{\sigma}^2)$与MLE相同。

4. 威布尔（Weibull）分布

$$V_1 = EX = \eta\Gamma(\frac{1}{m} + 1)，V_2 = \eta^2\Gamma(\frac{2}{m} + 1)，$$

故估计方程为：

$$\begin{cases} \eta\Gamma(\frac{1}{m} + 1) = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \eta^2\Gamma\left(\frac{2}{m} + 1\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}$$

与MLE不同。

第一节 参数估计的方法

5. 均匀分布 $X \sim U[0,\theta]$

此时$m=1$，令$EX=\frac{\theta}{2}=\bar{x}$，得$\tilde{\theta}=2\bar{x}$，

与MLE（$\hat{\theta}=\max\limits_{1\le i\le n} x_i$）不同。

哪个好，或各自优缺点?

● MLE（$\max\limits_{1\le i\le n} x_i$）概率为1地偏小。

● 矩估计在明知对某个$i\le n$，$\theta\ge x_i$（x_i较大）时，会仍用偏小的$2\bar{x}$作为估计。

第一节 参数估计的方法

三、一个实例

例1．"序列号"估计方法（背景：二战）

模型为N未知，从$\{1,2,\cdots,N\}$中随机抽取n个，记为$\{0<k_1<\cdots<k_n\leq N\}$，希望利用此数据估计$N$。

首先可以用$\frac{1}{n-1}\sum_{i=2}^{n}(k_i-k_{i-1}-1)$对两个数据间的平均间隔进行估计，从而$N$的一个自然的估计为：

$$\widehat{W}_1=k_n+\frac{1}{n-1}\sum_{i=2}^{n}(k_i-k_{i-1}-1)$$

$$=k_n+\frac{1}{n-1}(k_n-k_1)-1=\frac{n}{n-1}k_n-\frac{1}{n-1}k_1-1$$

可以证明，$E\left(\widehat{W}_1\right)=N$。

第一节 参数估计的方法

郑忠国提出了改进的估计：

$$\widehat{W}_2 = k_n + \frac{1}{n}\sum_{i=1}^{n}(k_i - k_{i-1} - 1)$$
$$= \frac{n+1}{n}k_n - 1$$

其中$k_0 =: 0$。亦可证明，$E\left(\widehat{W}_2\right) = N$，但其方差更小。

期望、方差表示什么？概率空间如何定义？

离散型均匀分布。

第二节 估计的优良性标准

当用$\varphi(x_1,\cdots,x_n)$估计目标$g(\theta)$时，希望$\varphi(x_1,\cdots,x_n)$与目标$g(\theta)$距离越近越好。但因为$\varphi(x_1,\cdots,x_n)$是随机变量，$g(\theta)$是θ的函数，也可以取不同的值，所以如何定义“距离”很重要。

在数理统计中，对同一随机事件，当参数θ取不同的值时，对应的概率测度也不同。为明确起见，记可测空间$(\Omega,\ \mathcal{F})$上的概率分布族为$\{P_\theta,\theta\in\Theta\}$，记相应的期望、方差等为$E_\theta()$、$Var_\theta()$等。

定义1. 称$\varphi(X_1,\cdots,X_n)$为$g(\theta)$的无偏估计，若

$$E_\theta\varphi(X_1,\cdots,X_n)=g(\theta),\qquad \forall\theta\in\Theta$$

这是较正常的要求。

第一节的常用分布中，两点分布p的MLE、正态分布中μ的MLE等都是无偏的，均匀分布的矩估计也是，但正态分布中 σ^2 的MLE不是无偏的！

第二节 估计的优良性标准

定义2. 设$\varphi(X_1,\cdots,X_n)$为$g(\theta)$的一个估计，称
$$M_\theta(\varphi)=E_\theta[\varphi(X_1,\cdots,X_n)-g(\theta)]^2$$
为φ的均方误差（Mean square error，MSE）。

若φ是无偏的，则$M_\theta(\varphi)=Var_\theta(\varphi)$。

定义3. 对于$g(\theta)$的两个估计φ_1，φ_2，若$\forall\theta\in\Theta$，有$M_\theta(\varphi_1)\le M_\theta(\varphi_2)$，则称$\varphi_1$不次于$\varphi_2$；若还存在 $\theta_0\in\Theta$，使得 $M_{\theta_0}(\varphi_1)<M_{\theta_0}(\varphi_2)$，则称$\varphi_1$比 φ_2有效。

例如，若 X的方差存在有限，θ是X的均值，令$\varphi_1(X_1,\cdots,X_n)=\frac{1}{n}\sum_{i=1}^n X_i$，$\varphi_2(X_1,\cdots,X_n)=\sum_{i=1}^n\lambda_i X_i$，其中$\lambda_i$满足$\sum_{i=1}^n\lambda_i=1$，则$\varphi_1$，$\varphi_2$均无偏；$\varphi_1$不次于$\varphi_2$；若$\lambda_i$不全相等，则$\varphi_1$比$\varphi_2$有效。

第二节 估计的优良性标准

●两个估计并不一定总能比较！一般地，不能找到不次于所有其他估计的估计量。（？）例如用常数$\varphi(X_1,\cdots,X_n)\equiv g(\theta_0)$估计$g(\theta)\cdots$。

●缩小范围，仅考虑无偏估计。

定义4. 称$\varphi(X_1,\cdots,X_n)$为$g(\theta)$的（一致）最小方差无偏估计（MVUE），如果它无偏，且对任意的$g(\theta)$的无偏估计$\psi(X_1,\cdots,X_n)$，有$M_\theta(\varphi)\le M_\theta(\psi)$（对$\forall\theta\in\Theta$）。（$\varphi$：\phi；$\psi$：\psi）

●最小方差无偏估计在许多常见情形下存在唯一，并可以求出。

●是一定标准下的“最优良”的估计。还有其他标准。

●如何求？可以通过利用充分统计量。

第二节 估计的优良性标准

定义5. 设$X_1,\cdots,X_n$为简单随机样本，称统计量$U=\varphi(X_1,\cdots,X_n)$是$\theta$的充分（sufficient）统计量，若似然函数$L(x_1,\cdots,x_n,\theta)$可表示为

$$q[\varphi(x_1,\cdots,x_n),\theta]\cdot h(x_1,\cdots,x_n)$$

其中$h(x_1,\cdots,x_n)$是不依赖于θ的非负函数。

●直观意义：充分统计量包含了样本$X_1,\cdots,X_n$中关于参数θ的全部信息。

●$\varphi_0(X_1,\cdots,X_n)=(X_1,\cdots,X_n)$一定是$\theta$的充分统计量。

●引入定义5的目的：降低样本的维数及复杂度时，不丢失关于θ的信息。

●等价表示：给定u_0及可测集A，$P_\theta((X_1,\cdots,X_n)\in A|U=u_0)$与$\theta$无关。

-----------------end 2024.02.22

有了充分统计量U，构造$g(\theta)$的估计时，可以仅考虑利用U即可。易知U的维数越低越好。